Multidimensional dialogue modelling

Petukhova, V.V.

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Multidimensional Dialogue Modelling

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
Tilburg University op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op donderdag 1 september 2011 om 16.15 uur door

Volha Viktarauna PETUKHOVA

geboren op 31 mei 1970 te Rybinsk, Sovjetunie

Promotor:

Prof. dr. H.C. Bunt

Samenstelling promotiecommissie:

Dr. J. Alexandersson
Prof. dr. A.P.J. van den Bosch
Prof. dr. N. Campbell
Dr. D.K.J Heylen
Prof. dr. E.J. Krahmer
Prof. dr. M.G.J Swerts
Dr. D. Traum

This research has been funded by the Netherlands Organisation for Scientific Research (NWO), under grant reference 017.003.090

TiCC Dissertation Series no. 17

Cover Design: Volha Petukhova using screenshots from the ‘Dimensions of Dialogue’ animation, created in 1982 by the Czech surrealist artist Jan Švankmajer

ISBN: 978-94-91211-88-1

Author email: v.v.petukhova@gmail.com

Acknowledgments

It has been a long way to this moment, but I enjoyed every mile of it. Obstacles made me only wiser, I gained a lot: knowledge, research experience, professional competence and confidence. I would like to thank many people who made the time during my PhD project stimulating and enjoyable. This thesis could not have been written without the help and inspiration of many people around me.

First of all, my special words of gratitude are to my promotor Prof. Harry Bunt for his immense support to this project from the very beginning of NWO application to the very end of finishing this thesis. His enthusiasm, interest and encouragement boosted my research many times. Harry, without your help this all would not be possible and this thesis would be twice as less valuable. Thank you also that you introduced me to the ISO world which gave me the unique opportunity to meet renowned researchers from all over the world, have frequent inspiring discussions with them, enhance my professional competence and broaden my knowledge.

My word of gratitude goes to the ISO editorial group members: Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, David Traum, James Pustejovsky, Martha Palmer and many others. It has been a great honor and a real pleasure collaborating with you. I would like to thank the members of the core PhD defense committee Jan Alexandersson, David Traum, Nick Campbell, Dirk Heylen, Antal van den Bosch, Marc Swerts and Emiel Krahmer for their valuable comments of the thesis.

An important part of the studies presented in this thesis is based on joint work with many colleagues. I would like to thank Jeroen Geertzen for constructive collaboration on dialogue segmentation and machine learning for dialogue act recognition, for being my room mate during the first three years and for comforting and supporting words during the hard phase of producing this thesis. I thank Laurent Prévot from the Laboratoire Parole et Langage of Université de Provence for a very pleasant and constructive work on discourse relations and hope our collaboration will not stop by this. Many thanks to Marcin Włodarczak from the Bielefeld University for work on ranking experiments and many insightful discussions. I would also like to thank my many Master students who participated in a lot of experiments reported in this thesis, but also for interesting papers on the Pragmatics course and Master thesis.

Many thanks to Mandy Schiffrin and Roser Morante, for their incredible support and that they agree standing by as paranymphs during my defense ceremony.

Many thanks go to my colleagues at the Tilburg Center for Cognition and Communication

for warm welcome in your research group, pleasant atmosphere and interaction: Peter Berck, Antal van den Bosch, Reinier Cozijn, Tanja Gaustad, Steve Hunt, Simon Keizer, Emiel Krahmer, Piroska Lendvai, Fons Maes, Roser Morante, Marie Nilsenova, Martin Reynaert, Marc Swerts, Paul Vogt, Menno van Zaanen, Kalliopi Zervanou, Lisette Mol, Herman Stehouwer, Sander Wubben, Marieke Hoetjes, Sander Canisius, Mandy Schiffrin, Ielka van der Sluis, and other people of the Faculty of Humanities. I would also like to thank many people from support staff, without their assistance and professionalism my stay at Tilburg University would be less pleasant and much less efficient: Joke Hellemons, Lauraine Sinay, Jacintha Buysse, Olga Houben, Lies Siemons, Peter van Balen, Leen Jacobs, and many others.

I would like to thank my teachers Elias Tijsse, Reinhard Muskens, Antal van den Bosch, Walter Daelemans, Emiel Krahmer, Marc Swerts, Ad Backus and Fons Maes for interesting lectures and for knowledge they shared with me during my two Masters at UvT.

At last, but certainly not at least, I would like to thank my family and friends for support and faith in me. I would like to thank my dearest partner Andrei for his encouragement, support, patience and understanding, but also as turned out for inspiring scientific discussions and cooperation. Your help has been decisive in the achievement of this goal!

Tilburg, *Volha Petukhova*
September 2011

Contents

Chapter 1

Introduction

1.1 Motivation

Multimodal natural-language based dialogue is increasingly becoming a feasible and attractive human-machine interface. Such interfaces offer a mode of interaction that has certain similarities with natural human communication, in using a range of input and output modalities which people normally employ in communication, such as speech, gesture, gaze direction and facial expressions. Some of these interfaces will advance to the incorporation of multimodality into virtual environments, for example as embodied conversational agents.

The design of dialogue systems that exhibit interactive behaviour which is natural to its users and that exploit the full potential of spoken and multimodal interaction may be expected to benefit from a good understanding of human dialogue behaviour and from the incorporation of mechanisms that are important in human dialogue.

Participation in dialogue is a complex activity in the sense that it involves not only the understanding and performance of actions for pursuing a certain goal or task; among other things, dialogue participants also constantly have to *"evaluate whether and how they can (and/or wish to) continue, perceive, understand and react to each other's intentions"* (Allwood, 1997). They share information about the processing of each other's messages, elicit feedback, manage the use of time, take turns, and monitor contact and attention. One of the reasons why people can communicate effectively and efficiently is because they use linguistic and nonverbal elements in order to address several of these aspects at the same time. Dialogue utterances, in other words, are often multifunctional. Consider, for example, the following dialogue fragment:[1]

(1) U1: Wat is RSI? /What is RSI?
S1: RSI staat voor Repetitive Strain Injury / RSI stands for Repetitive Strain Injury
U2: Ja maar wat is het? / Yes but what is it?
S2: Repetitive Strain Injury is een aandoening .../ Repetitive Strain Injury is an infliction ...

Utterance (U2) indicates that (1) the user understood the system's previous utterance (S1) (signalled by '*Ja/Yes*'); (2) the system did not interpret utterance (U1) as intended by the user (signalled by '*maar/but*'); and (3) the user requests information about the task domain. If the

[1] From a dialogue with the IMIX system - see Keizer and Bunt, 2007.

system does not recognize all three functions (and currently no system does), it will most likely resolve the anaphoric pronoun '*it*' as coreferential with '*RSI*' and interpret (U2) as a repetition of (U1), and thus not be able to react properly. This illustrates that the multifunctionality of utterances must be taken into account in order to avoid misunderstandings, and to support a dialogue that is effective and efficient.

Natural communication is also complex in the sense that dialogue participants use all available modalities in order to get their messages across. Face-to-face interaction involves besides speech also gestures, facial expressions, head orientation, posture, touch. A full-blown dialogue model has to take the contribution in each of these modalities into account, as well as their integration.

This thesis investigates some of the complexities of natural human dialogue by taking a multidimensional view on communication, and analysing dialogue behaviour as having communicative functions in several dimensions. Multidimensional approaches to dialogue analysis have been recognised to be empirically well motivated, and to allow accurate modelling of theoretical distinctions (Allwood (2000a), Allen and Core (1997), Bunt (1999), Klein (1999) and Larsson (1998)). Assigning communicative functions to utterances in multiple dimensions can help to represent the meaning of dialogue contributions in a more satisfactory way then is possible when only a single function is considered. Exploiting multidimensionality moreover supports a sensible segmentation of dialogue into meaningful units and improves system performance on the automatic recognition and interpretation of dialogue utterances.

The study presented in this thesis combines analytical and empirical investigations in order to build multidimensional computational dialogue models.

1.2 Research issues

Building a multidimensional dialogue model presupposes a clear and well-defined notion of 'dimension'. We will argue in some detail in Chapter 3 that the existing literature on multidimensional approaches to dialogue analysis does not provide such a notion. Multi*dimensionality* is often not clearly distinguished from multi*functionality*; an approach is often called 'multidimensional' if it supports the assignment of multiple communicative functions to dialogue utterances; the notion of dimension as such has not been analysed much. One of the first issues addressed in this thesis is how the notion of a dimension in the semantic and pragmatic analysis of dialogue can be defined, and what criteria can be used for identifying conceptually clear and useful dimensions. We will argue that the use of a well-defined notion of dimension leads to multidimensional approaches to dialogue analysis and dialogue modelling which are theoretically and empirically better motivated.

Since the multifunctionality of dialogue models is motivated in the first place by the multifunctionality of dialogue utterances, the notion of multifunctionality and its relation to 'dimensions' of communication deserves our attention. While it is widely acknowledged that dialogue utterances may have multiple communicative functions, there has hardly been any empirical study of this phenomenon. An issue that is addressed in this thesis is therefore which forms of multifunctionality are found in natural dialogue, and how these forms can be described and explained by taking a multidimensional perspective.

The assignment of communicative functions to stretches of dialogue and the forms of multifunctionality that are found when doing so, very much depend on two factors: (1) how these stretches of dialogue are chosen, and (2) whether only linguistically expressed functions are taken into account or also nonverbally expressed ones. The first of these factors is of crucial

importance, since longer stretches of dialogue obviously carry more communicative functions than shorter ones. Questions thus arise such as how a dialogue is best segmented into functionally meaningful segments, and how such segments can be defined and can be detected automatically. The second factor is equally important, since the use of nonverbal modalities such as head movements (e.g. nodding, shaking, waggling), gaze direction (e.g. looking at a dialogue partner; looking away), and facial expressions (e.g. smiling, frowning, blinking) gives a dialogue participant additional possibilities for expressing himself compared with the use of speech only. Does nonverbal behaviour in multimodal dialogue add to the (multi-)functionality of the interaction by introducing other functions than those that may be expressed linguistically in speech-only dialogue? This thesis address this question and more generally the multimodal expression and perception of communicative functions in dialogue.

For a dialogue system to be able to understand multifunctional utterances, it has to recognise utterance functions in context, and it has to do so on the basis of learnable features of utterances and dialogue context. Since people successfully interpret dialogue utterances incrementally, we want to explore to what extent and with what success rate we can simulate incremental segmentation and recognition of dialogue acts using available computational techniques. An utterance, when understood as a dialogue act with a certain communicative function and semantic content, evokes certain changes ('updates') in the context models of the dialogue participants. The formulation of an update semantics for multifunctional dialogue utterances calls for an articulate context model that enables multiple simultaneous and independent updates, and update mechanisms that describe how a participant's context model may change during a dialogue.

The studies in this thesis confirm that utterances in dialogue typically have multiple communicative functions. As a consequence, the utterances produced by a dialogue system will also be perceived by its users as having multiple functions. This is rather alarming, since existing dialogue systems do not generate utterances which are *meant* to be multifunctional, so this is a potential source of misunderstandings. This thesis explores the issue of how a dialogue system can generate utterances which are multifunctional by design, rather than by accident. Issues are addressed such as How can a Dialogue Manager generate multiple candidate dialogue acts, and What semantic, pragmatic, and empirical constraints should be taken into account when combining candidate acts for being jointly expressed in dialogue units of various sizes and forms.

1.3 Approach and starting points

The study presented in this thesis adopts an information-state or context-change approach (Poesio and Traum (1998); Bunt (1999); Larsson and Traum (2000)). This approach analyses dialogue utterances in terms of their effects on the dialogue contexts or 'information states' of participants. In particular, we use the theoretical framework of Dynamic Interpretation Theory (DIT) for its precise definitions of communicative functions and dialogue context.

Communicative functions are defined as specifications of the way semantic content is to be used by an addressee to update his information state when he understands the utterance correctly. This gives a formal semantics to the notions of communicative function and semantic content. We used the current version of the DIT dialogue act taxonomy, DIT^{++} Release 5 (see `http://dit.uvt.nl/`), which has been inspired by the goal to build a public registry of dialogue act specifications undertaken by the ISO organisation, and contains a well-defined set of dialogue act types with conceptually clear definitions.

Every communicative function is required to have some reflection in observable features of communicative behaviour, i.e. for every communicative function there are devices which a speaker can use in order to allow its successful recognition by the addressee. Such features may be linguistic cues, intonation properties, facial expressions, hand and head movements, etc. The analysis of the collected corpus data involved the identification of utterance features that can be used to detect the communicative functions of dialogue utterances (given certain context features), and in particular in order to investigate the automatic learnability of the communicative functions. The outcome of this part of our studies are the trained classifier(-s) to recognize multiple communicative functions on the basis of utterance and context features.

In DIT, a participant's dialogue context is understood as the totality of conditions that influence the generation and understanding of his dialogue behaviour. Dialogue acts are defined semantically as operators that update contexts in certain ways, which can be described by the communicative function and the semantic content of that dialogue act. The semantic content corresponds to what the utterance is about (what objects, events, etc., does it refer to; what propositions involving these elements are considered).

For developing a multidimensional model of dialogue context, we started from the DIT^{++} system of communicative functions and the DIT model of dialogue context, specifying them in more detail and representing the contents of context models by means of typed feature structures using the XML-based feature structure representation defined in ISO standard 24610-1; see Lee et al. (2004). The context model that was implemented in the PARADIME module of the IMIX dialogue system was taken as a starting point for this activity (Keizer and Bunt, 2006).

Our empirical studies of dialogue phenomena were supported by the analysis of empirical data collected in multimodal dialogue environments, in particular from the AMI and DIAMOND projects (see http://www.amiproject.org and http://pi1294.uvt.nl/diamond). Both speech and nonverbal behaviour in these dialogues were annotated in terms of dialogue acts, using existing annotation tools (notably ANVIL[2] and the DIT-tool[3]).

1.4 Contributions of this thesis

The contributions of this thesis fall into three categories: (1) fundamental concepts for dialogue modelling; (2) collection and analysis of multimodal dialogue data; (3) novel computational methods for dialogue analysis and context-driven dialogue management. We briefly indicate the main contributions in each of these areas.

Firstly, this thesis gives a definition of the notion of 'dimension' that has theoretical and empirical significance, and provides a basis for the choice of dimensions for multidimensional dialogue act taxonomies and annotation schemes. We formulated criteria that can be used to identify a dimension and to define a theoretically and empirically well-motivated set of dimensions. Application of these criteria led to the nine dimensions of the ISO 24617-2 dialogue act annotation scheme, and provided an underpinning for the set of ten dimensions in the DIT^{++} dialogue acts taxonomy.

Secondly, the multifunctionality of dialogue utterances is analysed. Where existing approaches define and study multifunctionality conceptually, almost exclusively taking theoretical considerations into account, the contribution of the thesis is that we investigate multi-

[2]For more information about the tool visit: `http://www.dfki.de/~kipp/anvil`
[3]For more information about the tool see Geertzen (2007).

functionality and its forms empirically as it is observed in dialogue data. For this purpose we collected and constructed multimodal dialogue data, which is itself a contribution in the second category. We developed the approach of multidimensional segmentation, and applied this method together with multidimensional annotation, showing (a) the feasibility of multidimensional segmentation applied to multimodal data; and (b) the applicability of multidimensional annotation schemes, developed primarily for spoken dialogue, to nonverbal and multimodal dialogue behaviour, provided that certain extensions are made for dealing with a speaker's uncertainty and sentiment.

A third contribution is the identification and successful application of features of nonverbal behaviour in the study of certain classes of dialogue acts, such as feedback acts, turn management acts, and discourse structuring acts. We revealed relations between observable features of communicative behaviour in different modalities and the intended multiple functions of multimodal utterances in dialogue. We also identified the general role of nonverbal signals for multimodal behaviour analysis in series of explorative and experimental studies.

In the area of computational dialogue modelling, a fourth contribution of this thesis is the development of a machine learning-based approach to the incremental understanding of dialogue utterances, with a focus on the recognition of their communicative functions. We combined local classifiers that operate on low-level utterance and context features with global classifiers that incorporate the outputs of local classifiers applied to previous and subsequent tokens. This approach resulted in excellent dialogue act recognition scores for unsegmented spoken dialogue. When a dialogue act is understood this evokes certain changes in the information states of the dialogue participants. Since we may deal with multiple simultaneous updates, due to the multiple communicative functions that an utterance may have, we specified a structured context model that enables multiple simultaneous and independent updates. We have outlined a context-driven approach to dialogue act interpretation and generation that enables the construction of intentionally multifunctional dialogue contributions. We studied dialogue act combinations empirically and analytically, and identified semantic, pragmatic, and empirical constraints that should be taken into account when combining candidate dialogue acts for producing multifunctional dialogue units of various sizes and forms.

The results described in this thesis can be profitably used for designing dialogue management tools as components of user-interface design in multimodal applications (such as embodied conversational agents), for the development of multidimensional annotation tools for multimodal dialogue, and for the automatic understanding and generation of (multifunctional) spoken or multimodal dialogue utterances. More generally the thesis contributes to the understanding of mechanisms in human dialogue, to the construction of annotated multimodal dialogue corpora, and to the development of dialogue systems that allow efficient and pleasant interaction with human users, exploiting the use of multiple modalities, of multifunctional contributions, and of rich context models.

1.5 Thesis outline

The thesis is organized in the following way.

Chapter 2 is concerned with theoretical and empirical aspects of dialogue analysis and dialogue modelling. Fundamental notions of dialogue theory are reviewed, the concept of dialogue act is introduced and some of its application, and alternative approaches to computational dialogue modelling are discussed.

Chapter 3 introduces the notion of dimension. We turn the readers' attention to the fact that the notions of 'dimension' that have been proposed in the literature are unsatisfactory in several respects. Dimensions are primarily used to group semantically similar communicative functions into one part of a dialogue act annotation scheme. We argue, however, that the notion of dimension has a conceptual, theoretical and empirical significance not only for annotation, but also for dialogue segmentation and interpretation, and enables more adequate dialogue modelling. Dimensions carry an essential part of the meaning of many dialogue utterances, and an adequate characterization of this aspect of meaning requires a coherent system of well-defined dimensions. We formulate requirements for distinguishing a dimension and for defining a coherent set of dimensions.

Chapter 4 addresses the dialogue act annotation task. Multidimensional and single-dimensional approaches to this task are discussed and compared. The semantic framework of Dynamic Interpretation Theory and in particular the DIT^{++} dialogue act taxonomy are introduced. Improvements and extensions are proposed. The annotation work is discussed that we performed, describing corpus data, transcriptions, and issues of dialogue segmentation. Basic concepts, a metamodel for dialogue act annotation that emerged from the collaborative research efforts within the ISO project 24617-2 'Semantic annotation framework – Part 2: Dialogue acts' are presented and elaborated.

Chapter 5 discusses forms of multifunctionality. Semantically different forms of multifunctionality are described and the actual co-occurrence of dialogue acts in different types of dialogue units is examined. The results of this study do not only have consequences for the semantic interpretation of dialogue contributions, but also for their generation by spoken dialogue systems.

Chapter 6 is concerned with the interpretation of communicative behaviour that it is observed in the annotated dialogue corpora. We focus on non-task related dialogue acts, mainly on feedback, turn management and discourse structuring mechanisms. We go into detail how single and multiple functions in these dimensions are expressed in different types of dialogue units, what linguistic and nonverbal means dialogue participants use for these purposes, and what aspects of a participant's behaviour are perceived as signals of these intentions.

Chapter 7 investigates automatic incremental dialogue act understanding using a token-based approach to utterance interpretation. We investigate the automatic recognisability of multiple communicative functions on the basis of the observable features such as linguistic cues, intonation properties and dialogue history. We show that a token-based approach combining the use of local classifiers, which exploit local utterance features, and global classifiers which use the outputs of local classifiers applied to previous and subsequent tokens, is shown to result in excellent dialogue act recognition scores for unsegmented spoken dialogue.

Chapter 8 outlines a context-driven approach to interpretation and generation of dialogue acts. We present a multidimensional context model and show how (multiple) dialogue acts correspond to (multiple) context update operations on this model. A formalization of dialogue act update effects is proposed. The context-based generation of dialogue acts is addressed as well as the selection of alternative admissible dialogue acts. We formulate semantic, pragmatic and linguistic constraints on dialogue act combinations for various types of dialogue unit.

Chapter 9 draws conclusions from the main findings of the thesis, and sketches perspectives for future research on the basis of our results.

Chapter 2

Dialogue and dialogue acts

This chapter introduces those aspects of dialogue analysis and dialogue modelling that are most important for this thesis. We provide an overview of the paradigms and formalisations that form the background for the analysis in subsequent chapters. The concept of a dialogue act is discussed. Approaches to dialogue act annotation, dialogue interpretation and generation, and computational dialogue modelling are reviewed.

Introduction

Dialogue is the most natural and basic form of language use. Very young children learn how to communicate with parents, playmates and others long before they learn to read and write. Ironic is the fact that we still do not have much explicit knowledge about how to adequately characterize the meaning of utterances in dialogue. This makes computational dialogue modelling a challenging task. Dialogue modelling involves a broad range of questions, such as: What is meaning in dialogue; What does it depend on; What mechanisms govern communicative behaviour in dialogue; How do dialogue participants transfer and process information; Why and how do they interpret, understand and react to each others' behaviour in the way they do. Computational dialogue modelling analyses these and related questions with computational means, and aims to cast potential answers in the form of computational models. The research presented in the following chapters addresses all these questions to some extent by applying a multidimensional, action-based analysis framework to the study of dialogue behaviour. This chapter mainly serves to provide the background for discussions and analyses presented later in this thesis.

We first discuss theoretical frameworks for dialogue analysis (Section 2.1). The aim is not to provide a historically complete overview of the various approaches, but rather to introduce and discuss the fundamental concepts that play a key role in this study. A discussion of the kinds of meaning that can be distinguished in dialogue brings us to the debate around the notion of 'dialogue act' (Section 2.2). Section 2.3 addresses the phenomenon of multifunctionality of dialogue utterances, that motivates the multidimensional analysis of natural human dialogue behaviour. Section 2.4 discusses the application of the dialogue act concept.

2.1 Dialogue theory

Central to a theory of dialogue is the understanding of dialogue behaviour. Bunt (1999) argues that a theory of dialogue cannot be expected to explain every word or turn in a dialogue, if only because the development of a dialogue often depends on properties of the task, for which the dialogue is intended to be instrumental, and which a theory of dialogue cannot reasonably be expected to take into account. Empirical studies of dialogue show however that dialogues exhibit regularities and patterns, both at the level of linguistic phenomena and other observable properties of communicative behaviour, and at the semantic-pragmatic level of communicative actions, and a theory of dialogue should be expected to interpret and explain the occurrence of such patterns and regularities.

Theoretical frameworks for dialogue analyses commonly assume that dialogue participants act as motivated, cooperative, rational and social agents (Clark, 1992; Clark, 1996; Sadek, 1991; Bunt, 1989; Bunt, 1999; Allwood, 1976; Allwood, 2000a; Allwood et al., 2000). Dialogue participants bring their own knowledge, beliefs, motivations, intentions and purposes; to communicate successfully, they have to coordinate their activities on many levels. They must share responsibilities for trying to solve problems (including communicative ones) to their collective satisfaction. Coordinating knowledge and beliefs is a central issue in all communications, and depends on the participants acting as motivated, cooperative, rational and social agents.

Motivation underlies any action, and often involves cooperation, ethics, power and esthetics (Allwood, 1976; 2000b). Dialogues are motivated by goals which are often non-communicative in nature, such as to solve a problem, or to take a decision. Such a motivation is often called a task that underlies the dialogue. Communication also involves performing a communicative task (Bunt, 1994): ensuring contact, providing feedback, monitoring attention, taking and giving turns, repairing communicative failures, and so on. Dialogue participants may adapt their personal goals to a common goal, but this is not always the case. Communicative agents may be motivated by their own goals, by their partner's goals, or by common goals.

Communication is always **cooperative** at some levels even if it involves conflicts. Communicative agents are cooperative at least in trying to recognise each other's goals, and the recognition of a goal may be sufficient reason for the participant to form the intention to act. Being a fully cooperative agent implies (Allwood et al., 2000):

1. to take each other into cognitive consideration: attempt to perceive and understand another person's actions, both communicative and not communicative;
2. to have a joint purpose (mutual contribution to shared purpose, mutual awareness of shared purpose, agreement made about purposes and antagonism involved in the purpose);
3. to take each other into ethical consideration (make it possible for others to act freely, help others to pursue his/her motives, make it possible for others to exercise rationality successfully);
4. to trust each other with regard to 1-3.

Rationality is analysed by Allwood (2000a) and Sadek (1991) in terms of adequate, efficient and competent action. People communicate with the aim to achieve something (underlying task) and they do this in a rational fashion (Bunt, 1994), organising the interaction so as to optimise the conditions for successful communication. A rational agent acts only if he thinks

it is possible to achieve an intended purpose (Allwood et al., 2000). People are capable of motivated action, and they often take each other's actions, motivations and other mental attitudes into consideration when acting. Each participant has functional as well as ethical tasks and obligations. The golden rule of ethics *'Do unto others what you would have them do unto you'* means in communication: *'make it possible for others to be rational, motivated agents'* (Allwood et al., 2000).

Dialogue communication is also a **social** activity. Communicative partners in dialogue act according to the norms and conventions for pleasant and comfortable interaction (Bunt, 1996). Communicative acts like greetings, apologies, and expressions of gratitude, agreement, or sympathy are often motivated by social obligations.

The assumption that dialogue participants perform cooperative, motivated, intentional, rational and social behaviour facilitates the understanding of phenomena and patterns in dialogue, discovering and explaining relations between communicative behaviour and participant goals, beliefs, preferences, and other aspects of mental states.

Communicative acts are often defined as acts with a conscious intention by the sender to transmit a certain message to the receiver. The question of the **conscious intentionality** of communicative acts deserves further discussion. An act which is not consciously intentional may still be relevant for analysis. For example, a lot of facial expressions are produced by humans unconsciously, but they display an emotional or cognitive state, which is obviously important for dialogue analysis. Goffman (1963) points out that the receiver is always responsible for the interpretation of an act as being intentional or not. Kendon (2004) also notices that whether an action is deemed to be intended or not is something that is dependent entirely upon how that action appears to others. This suggests that communication is a *joint activity* where the sender is responsible for encoding his intentions according to shared heuristics and expectations that makes it possible to interpret this behaviour, while the receiver is responsible for decoding the intended meaning by observing the sender's behaviour.

Allwood (1977) proposed criteria for the identification of communicative action. He argues that the identity of a communicative action should be determined in exactly the same way as the identity of any other action. He sees an action as a combination of:

- intention and purpose that an agent connects with an action;
- behavioural form an agent exhibits in performing an action (e.g. linguistic form);
- effects or results of a certain type of behaviour;
- context, because an action of a specific type occurs in a certain context.

Allwood (2000a) argues that each of these criteria can be a sufficient condition for saying that an action has occurred. He notices that communicative acts need neither necessarily be resultative nor intentional. An individual communicator can perform a communicative act without being perceived or understood (e.g. in a noisy environment, in a dialogue between participants who are non-native speakers with no or insufficient language skills); or can make a contribution unintentionally (e.g. this occurs often in the case of nonverbal acts); or a contribution does not need to be responded to and still will be counted as a communicative act leading to communication (e.g. beggars on the street, when their requests for money are ignored).

Allwood's criteria can be used to identify the type of action. For example, an Inform act could be characterised as follows:

- intention of performer: to provide the addressee with certain information in the form of a proposition *p*

- form of the behaviour: speaker utters a declarative sentence with content *p*
- achieved result: addressee believes that *p* is true
- context in which behaviour occurs: speaker and addressee are in contact, speaker believes *p* to be correct, speaker believes that addressee has no information about

One can produce an utterance of the form of an Inform when not all context conditions hold, e.g. when the addressee believes that *p* and the speaker is aware of this, but decides to remind the addressee. Or the form of the utterance could be different from a declarative sentence, e.g. rhetorical questions may be used as Informs.

Traum (2000) notices that using Allwood's criteria of communicative action can lead to misunderstandings among analysts and annotators as to whether a particular act has been performed, and whether the performance of an act implies a particular result. He argues that the kinds of conditions and their necessity may depend on the task being attempted. It also makes a difference whether this ascription is made from the point of view of an online dialogue participant or from that of an external observer, e.g. an annotator. Traum's remarks are very valuable. He points out that in formal dialogue theories actions are usually seen as transitions from one state to another, while dialogue acts are seen as special cases of actions. These theories describe dialogue acts as having an effect on the dialogue context, mental states, or social context. This is known in the literature as the *information-state* or *context-change* approach (Bunt, 1989; Traum and Larsson, 2003; Cooper et al.2003). These researchers generally associate several sets with actions: a set of effects (constraints on the resulting state), a set of pre-conditions (constraints on the initial state), and decompositions (sub-actions that, performed together constitute the action). In Allwood's terms, effects corresponds to achieved result, aspect(-s) of context and intention are related to the pre-conditions, and the form of behaviour is characterised by the decompositions. Traum notices that three aspects of context could be relevant for defining dialogue act types: dialogue state encoded in dialogue grammar (e.g. Traum and Allen, 1992; Lewin, 1998) or structural representation of context (e.g. Ginzburg, 1998); planning in terms of mental states of the speaker and addressee (*beliefs* and *intentions*, e.g. Allen and Perrault, 1980); and the third one in terms of the social obligations and commitments undertaken by the dialogue participants (e.g. Allwood, 1994). Most approaches combine two or three of these kinds of conditions and effects, for example, Dynamic Interpretation Theory (Bunt, 1989; 1994; 2000; 2005).

Dynamic Interpretation Theory (DIT) has emerged from the study of spoken human-human information dialogues, with the aim of uncovering fundamental principles to be applied in the design of human-computer dialogue systems. DIT models communicative agents as structures of goals, beliefs, preferences, expectations, and other types of information, plus memory and processing capabilities such as perception, reasoning, understanding, and planning. Part of these structures is dynamic in the sense of changing during a dialogue as the result of the agents perceiving and understanding each other's communicative behaviour, of reasoning with the outcomes of these processes, and of planning communicative and other acts (Bunt, 1999). DIT takes a context-change approach to dialogue acts and considers utterance meaning in terms of how they affect the context.

2.2 Dialogue acts

The notion of a dialogue act is a key notion in theories of dialogue. Dialogue acts are often used in studies of dialogue phenomena, in describing the interpretation of communicative behaviour

of participants in dialogue, and in the design of dialogue systems. Describing communicative behaviour in terms of dialogue acts is a way of characterizing the meaning of the behaviour. The idea of interpreting dialogue behaviour in terms of communicative actions such as statements, questions, promises, requests, and greetings, goes back to speech act theory (Austin, 1962; Searle, 1969), which has been an important source of inspiration for modern dialogue act theory.

Informally speaking, a dialogue act is an act of communicative behaviour performed for some purpose, e.g. acts provide information, request the performance of an action, apologise for a misunderstanding, and provide feedback. ISO standard 24617-2 defines a dialogue act as

(2) *communicative activity of a participant in dialogue, interpreted as having a certain communicative function and semantic content*[1]

A *communicative function* specifies the way semantic content is to be used by the addressee to update his context model when he understands the corresponding aspect of the meaning of a dialogue utterance.

In practice, two approaches can be found to defining communicative functions: (1) in terms of the effects on addressees intended by the sender; (2) in terms of properties of the signals that are used. Defining a communicative function by its linguistic form has the advantage that its recognition can be straightforward, but has to face the problem that the same linguistic form can often be used to express different communicative functions. For example, the utterance *Shall we start?* has the form of a question, and can be intended as such, but can also be used to invite or suggest somebody to start.

ISO standard 24617-2 takes a strictly semantic approach to the definition of communicative functions, but insists that for every communicative function there are ways in which a sender can indicate that his behaviour should be understood as having that particular function.

The second main component of a dialogue act is its *semantic content*, indicating what is the behaviour is about: which objects, events, situations, relations, properties, etc.

Semantically, dialogue acts can be viewed as corresponding to update operations on the information states of understanding participants in the dialogue (Bunt, 1989; Bunt, 2000; Traum & Larsson, 2003). For instance, when an addressee understands the utterance *Do you know what time it is?* as a question about the time, then the addressee's information state is updated to contain (among other things) the information that the speaker does not know what time it is and would like to know that. If, by contrast, an addressee understands that the speaker used the utterance to reproach the addressee for being late, then the addressee's information state is updated to include (among other things) the information that the speaker does know what time it is. Distinctions such as that between a question and a reproach concern the communicative function of a dialogue act.

2.3 Multifunctionality and multidimensionality

An utterance in dialogue may correspond to more than one dialogue act, and thus be multifunctional, for several reasons. Participation in a dialogue involves several activities beyond those strictly related to performing the task or activity. Dealing with the underlying task is very often combined in one utterance with pure communicative aspects such as the processing

[1] A note, added to the definition, remarks that "A dialogue act may additionally have certain functional dependence relations, rhetorical relations, and feedback dependence relations".

of each others messages, the use of time, taking turns, monitoring contact and attention. For example:

(3) 1. A: Do you know what date it is?
2. B: Today is the fifteenth.
3. A: Thank you.

In (3.3), A's utterance has the function of thanking, and will mostly be taken to imply that A has understood and accepted the information in (3.2) - i.e. as having a positive feedback function. But *'Thank you'* does not *always* express positive feedback; a speaker who finds himself in a rather unsuccessful dialogue may just want to terminate the interaction in a polite way. The feedback function of the thanking behaviour in example (3) can be inferred along the following lines: By saying *Thank you*, A thanks B, so there must be something that A is thankful for. This can only be what B just said, and that can only constitute a reason for thankfulness if A considers B's utterance as relevant and useful, which means that A accepted B's utterance as an answer to his question. The feedback function in such a case can be viewed as a conversational implicature (Grice, 1975).

There are also cases of multifunctionality where the different functions do not have any logical or implicature relations (see Chapter 5 for discussion of various forms of multifunctionality). This is for example the case for turn-initial hesitations, as in the following example:

(4) 1. A: Is that your opinion too, Bert?
2. B: Ehm,.. well,... I guess so.

In the first turn of (4), speaker A asks a question to B and assigns the turn to B (by the combined use of Bs name, the intonation, and by looking at B). In (4.2) B performs a stalling act in order to buy some time for deciding what to say; the fact that he starts speaking without waiting until he has made up his mind about his answer indicates that he accepts the turn. So the segment *Ehm,.. well,...* has both a stalling function and a turn-accepting function. Note, incidentally, that A's utterance is also multifunctional: it asks a question about B's opinion and it assigns the turn to B.

2.4 Use of dialogue acts

2.4.1 Dialogue annotation

According to the ISO Linguistic Annotation Framework (ISO 24612:2009) the term 'annotation' refers to the linguistic information that is added to segments of language data and/or nonverbal communicative behaviour. *Dialogue act annotation* is the activity of marking up stretches of dialogue with information about the dialogue acts performed, and is usually limited to marking up their communicative functions using a given set of such functions (a 'tag set').

Popescu-Belis (2005) identifies six types of constraints to be taken into consideration when designing a dialogue act tag set. A tag set should (1) relate to a theory of dialogue; (2) be compatible with the observed functions of actual utterances; (3) be empirically validated by high inter-annotator agreement (at least potentially); (4) facilitate automatic recognition of dialogue acts; (5) be designed with a particular NLP application in mind; and (6) be possible to map to existing tag sets.

Bunt (2005) emphasises that the annotation of dialogue corpus material brings specific constraints and requirements for a dialogue act annotation scheme, which should:

- support manual annotation, therefore definitions of dialogue act types and communicative functions should be in such terms that they facilitate human dialogue act recognition, and be clear enough to lead to consistent annotations with acceptable inter-annotator agreement;
- support automatic annotation, therefore dialogue act types and communicative functions should be defined in such terms as to facilitate the effective computation of dialogue act tags;
- support multidimensional annotation/interpretation: dimensions in a taxonomy should be independent as much as possible, and items within a dimension should be mutually exclusive except when they correspond to different levels of specificity;
- support different levels of granularity in annotations by reflecting different degrees of specificity in the (hierarchical) organisation of the taxonomy;
- use a terminology compliant with formal or de facto standards.

Another important part of an annotation scheme is *annotation guidelines*, which provide general principles and concrete instructions for how the tags should be used. They serve two main purposes: (1) to support the decision-making process of human annotators; and (2) to provide recommendations for possible extensions, modifications, or restrictions of the scheme as the need arises for particular applications.

Dialogue corpus annotation may serve various purposes. Annotated data is used for a systematic analysis of a variety of dialogue phenomena, such as turn-taking, feedback, and recurring structural patterns. Corpus data annotated with dialogue act information are also used to train machine learning algorithms for the automatic recognition and prediction of dialogue acts as a part of human-machine dialogue systems.

During the 1980s and 1990s a number of dialogue act annotation schemes have been developed, such as those of the TRAINS project in the US (Allen et al., 1994), the HCRC MapTask studies in the UK (Carletta et al., 1996), and the Verbmobil project in Germany (Alexandersson et al., 1998). These schemes were all designed for a specific purpose and a specific application domain. In the 1990s a general-purpose scheme for multidimensional dialogue act annotation was designed called DAMSL: Dialogue Act Markup using Several Layers (Allen and Core, 1997). Several variations and extensions of the DAMSL scheme have been constructed for special purposes, such as Switchboard-DAMSL (Jurafsky et al., 1997), COCONUT (Di Eugenio et al., 1998) and MRDA (Dhillon et al., 2004). The DIT^{++} scheme (Bunt, 2006 and 2009) combines the multidimensional DIT scheme developed earlier (Bunt, 1994) with concepts from DAMSL and various other schemes, and provides precise definitions for its communicative functions and dimensions. Chapter 3 discusses the most widely-used dialogue act annotation schemes and provides an overview of dimension-related concepts in these schemes.

2.4.2 Interpretation of dialogue behaviour

Interpretation of dialogue behaviour is primarily based on the recognition of the speaker's intentions. This raises the questions how dialogue participants signal their intentions, and what aspects of a participant's behaviour are perceived as signals of such intentions.

The state-of-art in dialogue act recognition is to use all available information sources from multiple modalities. These sources include: (1) linguistic information: lexical, collocational and syntactic features; (2) perceptual information including acoustic and prosodic properties of an utterance as well as information from visual and other modalities; (3) contextual information

that can be obtained from the preceding dialogue context as well as global context properties like dialogue setting, participant roles, and knowledge about dialogue participants.

The most studied dialogue act features are lexical cues. The presence or absence of particular lexical items in an utterance has for instance been used for identifying speaker intentions by Hirschberg and Litman (1993), Swerts and Ostendorf (1997), Jurafsky et al. (1998b) and Stolcke et al. (2000).

The role of prosody has been investigated by Shriberg et al. (1998); Jurafsky et al. (1998a); Lendvai et al. (2003); Swerts and Ostendorf (1997); Grosjean and Hirt (1996); Gravano et al. (2007); Hockey (1993); Nöth et al. (2002), to name few.

Another source of information for the interpretation of dialogue behaviour is knowledge of dialogue structure. Inspired by the observation that dialogue acts often come in so-called adjacency pairs (Schegloff, 1968), dialogue acts may be predicted from the occurrence of first elements of such pairs, see e.g. Nagata and Morimoto (1994); Woszczyna and Waibel (1994) and Stolcke et al. (2000).

In natural communication, the participants use all available modalities. This includes the use of gestures, facial expressions, gaze, posture shifts, speech and vocal sounds; communicative resources which make the communication richer in many ways. Visual cues for dialogue act interpretation have recently started to draw attention. Allwood (2000b) and Allwood and Cerrato (2003) emphasize the role of bodily communication for dialogue act interpretation in general, and for the interpretation of turn-taking behaviour and providing/eliciting feedback in particular. Cassell et a. (1999) and Cassell et al. (2001) study the role of gaze and posture shifts for discourse and information structure in dialogue. Kendon (2004) notices that some nonverbal acts can have various pragmatic functions: (1) a modal function, e.g. indicating whether the speaker regards what he is saying as a hypothesis or as an assertion; (2) a performative function, helping to indicate the kind of dialogue act, for example Offer - open palm-up hand movement; (3) a parsing function, e.g. punctuation, marking out logical components; (4) an interactive or interpersonal function, indicating focus of attention, attitude towards the addressee, social role in dialogue, right and obligations to occupy the sender role, and many others.

Chapter 6 will go into the details of how dialogue participants express the multiple functions of their contributions, and how they recognize the intended functionality of partner utterances. Chapter 7 will be concerned with the automatic recognition of dialogue acts based on features of natural human dialogue behaviour.

2.4.3 Dialogue models

In this section we discuss three prominent approaches to dialogue modelling: dialogue grammars, plan-based approaches, and the information-state paradigm.

Dialogue Grammar

Dialogue grammars are based on the observation that a dialogue exhibits certain regularities in terms of frequently occurring sequences of speech acts. For instance, questions are frequently followed by answers; requests and offers by acceptances or denials (Schegloff, 1968). Such *adjacency pairs* have been proposed to define grammar rules describing well-formed dialogues.

Examples of dialogue systems that use a dialogue grammar are SUNDIAL (Andry et al., 1990; Bilange, 1991) and LINLIN (Dahlbaeck and Jonsson, 1998).

	Request(Speaker,Hearer,Act)
CanDO.Pr	Hearer CanDo Act
Want.Pr	Speaker believe Speaker want$_{request-instance}$
Effect	Hearer believe Speaker want Act

Figure 2.1: Cohen and Perrault's definition of REQUEST.

The dialogue grammar approach has been criticized for being far from providing adequate explanation of dialogue behaviour. The model completely ignores (a) the semantic content of dialogue acts, and (b) the multifunctionality of dialogue utterances.

Plan-based approaches

Plan-based approaches to dialogue modelling are founded on the observation that participants in dialogue plan their actions to achieve certain goals. Allen (1983) argues that people are rational agents, forming and executing plans to achieve their goals, and inferring the plans of other agents from observing their actions. In order to understand what the speaker is saying an addressee uses both utterance properties and clues from his model of the speaker's cognitive state in order to recognise the plan that made the speaker say what he said.

While varying in their details, plan-based approaches (see e.g. Cohen and Perrault (1979), Allen and Perrault (1980), Sidner and Israel (1981), Carberry (1990) and Sadek (1991)) have in common that they view participating in dialogue in terms of speaker's **beliefs**, **desires** and **intentions**. Moreover, plan-based approaches relate a domain-level plan (e.g. a plan to get certain information, or to catch the train) with a communicative plan. Cohen and Perrault (1979) propose the use of formal plans that treat actions as operators, defined in terms of *preconditions*, *effects* that will be obtained when an action is performed, and a *body* that specifies the means by which the effects are achieved. Basically, they define two types of structures that a participant's mental state contains: *beliefs*, consists of an agent and a proposition which is believed by the agent, and *wants*, which represents the agent's goals. Figure 2.1 gives an example of how a Request is defined in terms of these operators.

Plan-based models assume a particular information flow for making inferences. First, a speech act is computed with its associated goal, then this information is used together with a domain plan to further specify the domain plan. A relationship between the current goal and the previous goal is constructed in order to infer implicatures of the current utterance, and therefore more information about the domain-level and communicative plans. This is what plan-based approaches are often criticized for. Plan construction and inference are activities that can easily get very complex and become computationally intractable. Moreover, some dialogue phenomena like actions that are not about planning or about the task at all (such as feedback, clarification questions, confirmations, etc., which constitute a great portion of all utterances in dialogue, see Chapter 3 of this thesis,) are difficult to model by means of plan recognition and plan generation. In order to overcome these shortcomings Grosz and Sidner (1986) and Grosz and Sidner (1990) proposed to consider conversation as a joint activity. According to this approach (known in literature as the *collaborative approach*) all dialogue partners work together to achieve and maintain understanding in dialogue. Collaborative approaches try to capture the motivation behind a dialogue and the mechanisms of dialogue itself, rather than focus on the structure of the task. This suggests that the beliefs of all dialogue parties should be modelled and if the proposed goal is accepted by another partner it will become part of the

shared (mutual) beliefs (see also Traum, 1994 and Traum, 1999).

Plan-based models have been applied for example in the TRAINS system (Allen et al., 1994) and in the TRIPS system (Allen et al. (2001), which has a task manager that relies on planning and plan recognition.

ViewGen (Wilks and Balim, 1991) is a system for modelling agents, their beliefs and their goals as part of a dialogue system, which uses a planner to simulate other agents' plans. Nested beliefs (about beliefs and goals) are created only when required as the plan is generated and are not pre-stored in advance before the plan is constructed, as in (Cohen and Perrault, 1979) and (Allen and Perrault, 1980).

The Verbmobil speech-to-speech translation system uses a plan recognizer similar to that of plan-based models (Wahlster, 2000).

The major accomplishment of plan-based theories of dialogue is that they offer a generalization in which dialogue can be treated as a special case of rational behaviour. The primary elements are accounts of planning and plan-recognition, employing inference rules, action definitions, models of the mental states of the participants, and expectations of likely goals and actions in the context. The set of actions may include dialogue acts, whose execution affects the beliefs, goals, commitments, and intentions of the conversational partners.

Information-state approaches

Information state update approaches, see Poesio and Traum (1998); Traum et al. (1999); Bunt (1989; 2000); Larsson and Traum (2000), analyse dialogue utterances in terms of effects on the information states of the dialogue participants. An 'information state' (also called 'context') is the totality of a dialogue participant's beliefs, assumptions, expectations, goals, preferences and other attitudes that may influence the participant's interpretation and generation of communicative behaviour (Bunt et al., 2010). Dialogue acts are viewed as corresponding to update operations on the information states of understanding participants in the dialogue.

An assumption that is shared between all proposals for information states (e.g. Poesio and Traum, 1998; Bunt,2000; Ahn, 2001; Cooper, 2004) is that an information state is structured into a number of distinct components. The information is for example divided into a 'private' part which contains *beliefs* which the participant assumes to be true; an *agenda* which contains short term goals or obligations of the agent; and a *plan* which contains actions or dialogue acts that the agent intends to carry out. A private part may also include 'temporary' shared information that has not yet been grounded, for instance including set of propositions that the participant *believes* to be true, a stack of *questions under discussion*, questions that have not been answered yet (see Ginzburg, 1998), and *latest utterance*, containing information about the latest utterance. The 'shared' part contains the same components as a 'temporary' shared one with the difference that this information has been grounded in dialogue, i.e. acknowledged by other participants. Figure 2.2 represents the information state of a dialogue participant as defined in (Traum et al., 1999).

Several dialogue system have been developed using such a framework, such as GoDIS (Larsson et al., 2000), IBiS1 (Larsson, 2002) and DIPPER (Bos et al., 2003).

2.5 Summary

In this chapter three main issues in dialogue analysis and dialogue modelling have been reviewed: the kind of principles that govern dialogue behaviour; the notion of utterance meaning

$$\left[\begin{array}{l} PRIVATE: \left[\begin{array}{l} PLAN: StackSet(Action) \\ AGENDA: Stack(Action) \\ BEL: Set(Prop) \\ TMP: \left[\begin{array}{l} BEL: Set(Prop) \\ QUD: Stack(Questions) \\ LU: \left[\begin{array}{l} SPEAKER: PARTICIPANT \\ MOVES: assocSet(Move, Bool) \end{array}\right] \end{array}\right] \end{array}\right] \\ SHARED: \left[\begin{array}{l} BEL: Set(Prop) \\ QUD: Stack(Questions) \\ LU: \left[\begin{array}{l} SPEAKER: PARTICIPANT \\ MOVES: assocSet(Move, Bool) \end{array}\right] \end{array}\right] \end{array}\right]$$

Figure 2.2: Example of information state as defined in Traum et al. (1999).

in dialogue; and the use of dialogue acts.

It has been observed by many researchers that human dialogue behaviour exhibits certain patterns and regularities. The assumption that dialogue participants act as motivated, cooperative, rational and social agents allows to find and explain such regularities and is extremely useful to model the fundamental aspects of dialogue communication. More specifically, the use of particular communicative acts in order to signal the speaker's state of beliefs, disbeliefs, and other attitudes, is governed by general principles allowing the interpreter to reconstruct the relevant aspects of the speaker's cognitive state. These principles and their application in the interpretation and generation of specific kinds of communicative act form a basis for constructing and updating articulate dialogue models.

The use of language (in a broad sense, including body language) in dialogue can be characterised in terms of communicative acts. It was noted that a communicative act can be defined using three main concepts: intention (or purpose), effects and context. A communicative act has a purpose and has certain effects on the addressee. The interpretation of intention and effects is context-dependent. Adequate characterization and formalization of communicative act semantics in terms of intended context-changing effects on participants' information state is an important step forward in the analysis of dialogue phenomena, in the description of the interpretation of communicative behaviour of dialogue participants, and in the design of dialogue systems. Such a characterization and formalization is provided by the notion of a 'dialogue act' (Bunt, 1989) seen as an update operator on information states, and having two main components: communicative function and semantic content. Thus, describing communicative behaviour in terms of dialogue acts is a way of characterizing the meaning of the dialogue behaviour, and the ultimate goal is to reconstruct the agent's intentions from the observation of his behaviour.

A phenomenon of fundamental importance is that dialogue contributions are often multifunctional. This has to be taken into account when modelling dialogue behaviour. DIT provides a framework for adequately characterising multifunctionality in terms of multiple dialogue acts performed simultaneously, addressing different independent communicative dimensions.

Three main uses of dialogue acts have been discussed: analysis of dialogue phenomena, dialogue annotation, and dialogue modelling. We considered the most widely used strategies, techniques and trends for analysing dialogue interaction. The dialogue act annotation task was outlined, and existing approaches to dialogue act annotation were brought up. Three prominent approaches to dialogue modelling were reviewed that make use of the notion of dialogue act: dialogue grammars, plan-based approaches and the information-state update framework. The

latter two, also in combination with other approaches, e.g. agent-based methods, allow for richer and more flexible dialogue modelling.

Chapter 3

Dimensions in dialogue interaction

This chapter provides a theoretical and empirical basis for the choice of dimensions in a multidimensional dialogue annotation and interpretation system. A 'dimension' in this context is a class of semantically related dialogue acts which has a proven conceptual and empirical significance. Five criteria are put forward which a set of such dimensions should meet: theoretical justification, empirical validity, orthogonality, *reliable* recognisability, *and* compatibility *with existing annotation schemes where possible. Applying a range of tests to annotated dialogue corpora, and taking 18 existing annotation schemes into account, ten dimensions are identified which are shown to meet these criteria.*

Introduction

The observation that dialogue behaviour is often multifunctional, in the sense of having more than one communicative function simultaneously, is partly explained by the fact that dialogue contributions may contain several functionally relevant stretches of behaviour. Even if minimal stretches are considered, such as one-token segments, multifunctionality does not go away. This phenomenon can be accounted for taking a multidimensional view on communication and analysing dialogue behaviour as having communicative functions in several dimensions. Dimensions are mainly concerned with dialogue underlying task or activity and purely communicative tasks, such as social obligations, structuring the discourse, managing contact, editing their own and partner's speech, etc. A set of dimensions that are theoretically and empirically justified can be a good foundation for a multidimensional dialogue act annotation scheme which can be used for an adequate analysis of human dialogue behaviour.

A variety of approaches can be found which make use of a notion of 'dimension'. In the 1990s a group of researchers came together as the Discourse Research Initiative and drafted the multidimensional dialogue act annotation scheme called DAMSL: Dialogue Act Markup using Several Layers (Allen and Core, 1997 and Core and Allen, 1997). DAMSL defines

Here and in the chapters 4-8 I describe in a chapter-initial note publications that the chapter is based on and the division of work, since I have published exclusively in collaboration with others. I would like to stress that this thesis constitutes original work and no chapter or part of it is based entirely on any one article. This chapter - to some extent adapted from the TiCC report by Petukhova and Bunt (2009) - is written by me, with comments, additions and proof by Harry Bunt.

four so-called layers: Communicative Status, Information Level, Forward-Looking Function (FLF) and Backward-Looking Function (BLF); the last two are concerned with communicative functions. The FLF layer is subdivided into five classes, including the classes of commissive and directive functions, well known from speech act theory. The BLF layer has four classes: Agreement, Understanding, Answer, and Information Relation. Core and Allen (1997) refer to these nine classes as 'dimensions'.

Soria and Pirrelli (2003) proposed a meta-scheme for comparing schemes along orthogonal 'dimensions' of analysis which have a bearing on the definition of dialogue acts. The following classificatory dimensions are defined: (D1) grammatical information; (D2) information about lexical and semantic content; (D3) co-textual information; (D4) pragmatic information. Comparing annotation schemes via a meta-scheme may enable a judgment of their similarity. Using such a meta-scheme for designing a comprehensive dialogue act scheme seems difficult and complicated, however. For example, an utterance like *'What time would engine two leave Elmira?'* would have the following annotation: (D1) wh-question; (D2) request-info; (D3) initiative and (D4) directive. This obviously contains a great deal of redundancy.

Popescu-Belis (2004) argues that dialogue act tag sets should seek a multidimensional theoretical grounding and defines the following aspects of utterance function that could be relevant for choosing dimensions in a multidimensional scheme: (1) the traditional clustering of illocutionary forces in speech act theory into Representatives, Commissives, Directives, Expressives and Declarations; (2) turn management; (3) adjacency pairs; (4) topical organization in conversation; (5) politeness functions; and (6) rhetorical roles. He observes that an utterance often has a function in several dimensions: for instance, every utterance also plays a role in turn management. Therefore, when looking for utterance functions, several dimensions should be considered. He proposed a tag set called 'Principled Multifunctional Annotation of utterances in dialog' (PRIMULA). It is however not obvious why the proposed six dimensions are chosen.

Several questions emerge from these proposals: (1) What is a 'dimension'? (2) Is there a concept of 'dimension' in the literature that we can use? and (3) What criteria can be established for distinguishing a 'dimension' in a multidimensional dialogue act annotation scheme? (4) When apply a sensible set of criteria, what dimensions do we get? This chapter is devoted to finding answers to these questions.

3.1 The notion of 'dimension'

As noted in the previous section, a variety of approaches can be found which make use of a notion of 'dimension'. A dimension is often conceived as a cluster of dialogue acts which are in some respect similar and which form a set of mutually exclusive tags that can be assigned independently from the tags in other dimensions as defined (e.g. Larsson, 1998). Such a definition is unsatisfactory in several respects. First, the functions that form a dimension do not need to be mutually exclusive. For example, the DAMSL dimension of Understanding has three functions: signal-non-understanding, signal-understanding, and correct-misspeaking. Of these, correct-misspeaking implies signal- understanding, because in order to make a correction the speaker has to understand the utterance which he believes to contain an error; hence these tags are not mutually exclusive.

Second, not every similarity relation is suitable for defining a dimension. For instance, similarity based on the type of communicative function is not satisfactory. For example, the cluster of 'information-seeking functions' for a range of question types, and the cluster of 'information-providing functions' for various kinds of informs, could be considered as dimen-

sions, as is the case in DAMSL. This would mean that an utterance may be tagged as being both a question and an answer concerning the same content. This seems highly undesirable.

In DAMSL a dimension is defined as *"an abstract characterisation of the content of an utterance"* (Allen and Core, 1997). It is noticed that *"in task-oriented dialogues, we can roughly divide utterances into those that address the task in some way, those that address the communication process (Communication Management), and those that do not fall neatly into either category (Other-Level). In addition, we can subdivide the first category into utterances that advance the task (Task) and those that discuss the problem solving process or experimental scenario (Task Management)"* (Allen and Core, 1997). This is a coarse distinction of semantic content types, which may be refined by subdividing Communication Management into feedback, turn management, topic management, and other aspects.

Bunt and Girard (2005) suggest that a well-founded notion of *dimension* can be based on the observation that participants in a dialogue are not only concerned with performing the task that underlies the dialogue, but also share information about the processing of each other's messages, about the allocation of turns, about contact and attention, and about various other aspects of interaction. They thus perform various types of communicative acts, such as giving and eliciting feedback, taking turns, stalling for time, establishing contact, and showing attention. Each of these types of communicative activity is concerned with a particular type of information: feedback acts are concerned with the success of processing previous utterances; turn management acts with the allocation of the speaker role; topic management acts with the topical progression of the dialogue, and so on. These observations lead to the following definition of the notion of a dimension:

(5) A dimension is a class of dialogue acts concerned with one particular aspect of communication, corresponding to a particular type of semantic content, which a dialogue act can address independently from other dimensions.

Dimensions, in the sense introduced here, classify dialogue acts. What is usually called a 'dialogue act taxonomy' is in fact a taxonomy of the *communicative functions* of dialogue acts (like Question, Offer, Confirmation, Signal-Understanding, Turn-Grabbing, Greeting, Stalling,...). Some communicative functions are always concerned with the same type of information, such as a Turn Grabbing act, which is concerned with the allocation of the speaker role, or a Stalling act, which is concerned with the timing of utterance production. Being specific for a particular dimension, such functions are called *dimension-specific*.

Other functions are not specifically related to any dimension in particular, e.g. one can ask a question about any type of semantic content, provide an answer about any type of content, or request the performance of any type of action (such as *Could you please close the door* or *Could you please repeat that*). These communicative functions are called *general-purpose* functions, and include Question, Answer, Request, Offer, Inform, and many other familiar core speech act types. Given a set of dimensions, the dialogue act that results from applying such a function to a particular content can be classified depending on the type of its content. Example (6) illustrates this for the Inform function.

(6) 1. I didn't hear what you said [*dimension: Feedback*]
2. I would like Peter to continue [*dimension: Turn Management*]
3. The next meeting will be on Friday December 3 [*dimension: Task*]
4. It'll take me a while to gather that information [*dimension: Time Management*]

3.2 Criteria for dimension identification

According to the definition of 'dimension' provided in (5) we need to identify relevant communicative aspects that dialogue acts are concerned with. This can be established both empirically and theoretically. Only dimensions should be considered which can be distinguished according to empirically observable behaviour in dialogue. Second, each dimension should be theoretically justified, i.e. corresponding to a well-established class of communicative activities, such as taking turns, monitoring contact and attention, and providing and eliciting feedback. A third criterion is that each dimension should be recognizable with acceptable precision by human analysts, as well as by dialogue understanding and dialogue annotation systems, in order to be useful for annotation and system design.

In addition to these three criteria, that apply to each individual dimension to be distinguished, a fourth criterion concerns the inclusion of a dimension in a *set of dimensions*: different dimensions should be concerned with clearly *different* aspects. More specifically, the dimensions in a multidimensional framework should be *'orthogonal'* or *independent*, in the sense that each of the dimensions can be addressed by dialogue acts independent from addressing other dimensions. This is a criterion that can be tested empirically.

Finally, a set of dimensions that is applicable for a wide range of task domains and types of communicative situation is evidently more valuable than one that has limited applicability. A proposed set of dimensions can be evaluated in this respect by considering dimensions that play a role in existing annotation schemes. This rather practical consideration can be turned into a criterion for including a dimension in a set of dimensions, namely that this dimension is found in a significant number of existing annotation schemes.

In sum, the following criteria can help to make a well-motivated choice of the dimensions in a general-purpose multidimensional dialogue act annotation scheme:

(7)
1. each dimension is theoretically well established;
2. each dimension is empirically observed in the functions of dialogue utterances;
3. each dimension is recognizable by human annotators and by automatic systems;
4. each dimension is orthogonal to all other dimensions in the set of dimensions;
5. each dimension is found in a significant number of existing annotation schemes.

Using these criteria, the study reported in this chapter aims to provide a theoretical and empirical basis for choosing dimensions in a multidimensional scheme.

3.3 Methodology

We applied the criteria listed in (7) in a partly theoretical and partly empirical study. This study benefited from surveys that were conducted in the EU-funded projects MATE[1] (Klein and Soria, 1998) and LIRICS (Bunt and Schiffrin, 2007). In the MATE project, carried out in 1996-1999, 16 dialogue annotation schemes were analysed. However, some of the schemes developed around this time were not considered, such as the second revised version of Verbmobil (Alexandersson et al., 1998).

[1] **M**ulti level **A**nnotation **T**ools **E**ngineering (`http://www.ims.uni-stuttgart.de/projekte/mate/`)

More recently, several new schemes have been developed which are analysed in this study. The multidimensional MRDA scheme (Dhillon et al., 2004) was developed for the purpose of analysing conversations in meetings. In the AMI project,[2] which aimed to develop technologies for meeting support, a dialogue act annotation scheme was developed (AMI Consortium, 2005b) and a 100-hours meeting corpus was annotated using this scheme. Although one-dimensional, the AMI scheme has some features that allow more accurate dialogue act annotation than other one-dimensional schemes. An additional layer of so-called 'reflexive' acts allows labelling the type of semantic content by specifying whether a dialogue contribution is about the meeting task or about managing the task.

The DIT^{++} annotation scheme (Bunt, 2006; Bunt, 2009a) combines the multidimensional DIT scheme developed earlier (Bunt, 1994) with concepts from DAMSL and various new schemes, and provides precise definitions for its communicative functions and dimensions. DIT^{++} release 2 contained 11 dimensions, including a dimension concerned with monitoring and managing difficulties in the partner's contribution and called *Partner Communication Management*.[3] The most recent Release 5 (from May 2010) contains 10 dimensions; Topic Management and Discourse Structuring were merged into one Discourse Structuring dimension.

In 2006 the LIRICS project was launched as a preparatory step for an ISO project aiming to develop annotation standards. A set of core communicative functions from the DIT^{++} scheme was redefined using ISO standard 12620 for the specification of data categories; these were tested for their usability in manual dialogue annotation in English, Dutch and Italian (Bunt and Schiffrin, 2007). It was decided to merge the dimensions of *Discourse Structuring* and *Topic Management* since they are not orthogonal. The resulting set of 10 dimensions was taken as a starting point for this study.

The criteria listed in (7) can help us to obtain a clear picture of which semantic clusters in the various multidimensional schemes might qualify as dimensions. For instance, of the groupings defined in DAMSL (Allen and Core, 1997) Task, Task Management, Communication Management and Understanding are good candidates, since they are theoretically distinguished and independent aspects of communication, while Info-Request, Statement, Directives and Commissives are not, as argued above. From the dimensions proposed by Popescu-Belis for the PRIMULA tag set (Popescu-Belis, 2004), Turn Management, Topic Organisation and Politeness do qualify as potential dimensions. Dimensions proposed by Allwood and colleagues for the SLSA annotation scheme (see Allwood et al., 1997; Allwood et al., 1993; Nivre et al., 1998) include Feedback, Turn Management and Own Communication Management. The LIRICS scheme includes the dimensions of Partner Communication Management, Contact Management, Social Obligation Management, Discourse Structuring (slightly broader then Popescu-Belis' 'Topic Organisation') and Time Management.

For testing the criterion of empirical validity relating to communicative dimensions we analysed the following dialogue corpora:

- the DIAMOND corpus[4], which consists of two-party human-human instructional task-oriented dialogues in Dutch;
- the AMI meeting corpus, which consists of multimodal task-oriented human-human multi-party dialogues in English;

[2] **A**ugmented **M**ulti-party **I**nteraction (`http://www.amiproject.org/`)

[3] Release 2 is from October 2006; see `http://dit.uvt.nl/`.

[4] For more information see Geertzen,J., Girard,Y., and Morante,R. 2004. The DIAMOND project. Poster at the 8th Workshop on the Semantics and Pragmatics of Dialogue (CATALOG 2004).

- the OVIS corpus[5], which consists of task-oriented human-computer dialogues over the telephone in Dutch.

All corpora were manually segmented into functional segments[6] and tagged using the LIRICS annotation scheme. The DIAMOND dialogues contain 1,408 functional segments; the AMI dialogues 3,897; and the OVIS corpus 3,942. We analysed the distribution of the tags that were used in the various dimensions (Section 3.5), and conducted a series of recognition experiments (Section 3.6).

The orthogonality of a set of dimensions can be determined empirically and theoretically. Theoretically, dependency relations can be uncovered by analysing the definitions of dimensions and their function tags, in particular for the existence of logical relations between the preconditions of communicative functions. For example, a *dialogue opening* is logically related to a *contact indication* act, because the precondition for a contact indication act, which says that the speaker wants the addressee to know that the speaker is ready to communicate with the addressee, is among the preconditions of a dialogue opening (see Chapter 8, Section 8.2).

Empirically, dependency relations can be found by analysing annotated dialogue data. Tags which always co-occur are either logically related or else show an empirical fact about communication; similarly for zero co-occurrence scores. Besides co-occurrence scores, we also provide a statistical analysis using the phi coefficient as a measure of relatedness. The phi measure is related to the chi-square statistic, used to test the independence of categorical variables. In addition, to investigate whether dimensions are concerned with very different information, we defined the similarities between dimensions in terms of distances between dimension vectors in a multidimensional space, where orthogonal vectors convey unique, non-overlapping information.

If a dimension is not independent from other dimensions, then there would be no segments in the data which address only that dimension. Looking for segments which address *only* one dimension is therefore another test. Finally, we investigate whether a dimension is addressed always in reaction to addressing a certain other dimension. If that is the case, then the presence of a dimension in a multidimensional scheme depends on the presence of another dimension. For example, the *answer* dimension as defined in DAMSL cannot be seen as an independent dimension because *answers* need *questions* in order to exist. The test here is to examine for each dimension the relative frequencies of pairs <dimension tag, previous dimension tag>; if a tag always co-occurs with a certain previous tag, then there is apparently a dependence between the two.

To sum up, we performed five tests, examining:

1. the relative frequency of *communicative function co-occurrences* across dimensions;
2. *the extent of relatedness between dimensions* measured with the phi coefficient;
3. *dimension vector distances* in multidimensional space;
4. for all dimensions whether there are functional segments *addressing only that dimension*;
5. the relative frequency of pairs of *dimension* and *previous dimension*.

Dependency tests are reported in Section 3.7.

[5] **O**penbaar **V**ervoer **I**nformatie **S**ystem (Public Transport Information System) `http://www.let.rug.nl/~vannoord/Ovis/`

[6] Functional segments are defined as a minimal stretch of communicative behaviour that has a communicative function (and possibly more than one), see Geertzen et al., 2007.

3.4 Theoretical validation

Dialogue purpose and domain of discourse

Dialogues are a form of motivated rational behaviour (Allwood, 2000a), inspired by goals, tasks, or activities which are non-communicative in nature, e.g. aiming to obtain certain information, to get someone's support, to improve relationships, or to play a game. We will use the term *task* for referring to this underlying activity or motivation. Dialogue participants are assumed to pursue a certain task in a rational fashion, organising the interaction so as to optimise the conditions for successful communication.

Contact, presence, and attention

Communication presupposes that the parties are in *contact* and are willing to be in continued contact. *"If A attempts to communicate with B, he/she can expect B to respond, at least by indicating that no contact is possible, and any response from B is enough to manifest contact"* (Allwood et al., 2000). This aspect of communication is of a particular importance if there is no or limited visual contact between the participants. But also when there is direct visual contact, the participants tend to permanently check the attention of their interlocutors and their readiness to continue the conversation. Body movements and facial expressions (e.g. gaze direction) are used for this purpose (Goodwin, 1981).

Grounding and feedback

For successful communication, dialogue participants have to coordinate their activities on many levels other than that of the underlying task. The coordination of knowledge and beliefs is a central issue in communication; Clark (1996) argues that speakers and addressees attempt to establish the mutual belief that the addressee has understood what is uttered. The process of establishing mutual understanding of each other's intentions and actions is known as *grounding*. Traum (1999) proposes to distinguish a class of grounding acts; which are closely related to *feedback*. Feedback mechanisms, their linguistics (verbal and non-verbal expressions, durational, temporal and prosodic properties) and related phenomena have been studied extensively, e.g. Duncan and Fiske (1977); Allwood et al. (1993); Clark and Krych (2004). Allwood et al. (1993), Clark (1996) and Bunt (2000) distinguish several *feedback levels*: attention (called contact in Allwood et al., 1993), perception (identification in Clark, 1996), understanding (interpretation in (Bunt, 2000), evaluation (consideration in (Clark, 1996) and attitudinal reaction in (Allwood et al., 1993)), and execution, defined in (Bunt, 2000).

A speaker may provide feedback on his own processing of previous utterances (*feedback giving* functions or auto-feedback, in terms of Bunt (1994)), or elicit feedback when he wants to know the processing status of the addressee (*feedback eliciting* functions, or provide feedback on the partner processing of previous utterances, called *allo-feedback*, in the terminology of Bunt (1994)).

Taking turns

Allwood (2000a) defines turn management as the distribution of the right to occupy the sender role in dialogue. He argues that a turn is a normative rather than a behavioural unit. In the well-known study of Sacks, Schegloff and Jefferson (1974) it was observed that in a wide

range of contexts most of the time only one of the participants in the conversation is talking; that occurrences of more than one speaker at a time were brief; and that transitions from one turn to the next without a gap or overlap were very common.

Recent years have seen a number of qualitative and quantitative findings on turn-taking mechanisms and related phenomena (e.g. Cassell et al., 1999, Selting, 2000, ten Bosch et al., 2004, Campbell, 2008) studying the ways in which dialogue participants indicate that they intend to start speaking, finish speaking, resume speaking, or give the right to speak to someone else.

Social obligations and politeness

Participating in a dialogue is a social activity, where one is supposed to do certain things and not to do others, and to act in accordance with the norms and conventions for social behaviour. A dialogue participant has besides functional also ethical tasks and obligations, and performs social obligation management acts to fulfill these. Such acts are closely related to politeness phenomena (see e.g. Lakoff (1973) and Brown and Levinson (1987)).

Bunt (1996) noticed that many social acts are not just 'social', they also improve the transparency of the dialogue. For example, people greet each other also for establishing their presence, and say good-bye also to close the conversation.

Dialogue structure

A speaker may indicate his view of the state of the dialogue, and makes the hearer acquainted with his plans for dialogue continuation, e.g. that he is going to close the discussion of a certain topic; or that he wants to focus the hearer's attention on a new topic.

The organization of discourse has been studied extensively, e.g. for monologues by Mann and Thompson (1988); and for dialogues by Asher and Lascarides (2003), Hirschberg and Litman (1993), and Heeman and Allen (1999) among others. A distinction is sometimes made between *macro-*, *meso-* and *micro-levels* in discourse structuring (e.g. Nakatani and Traum, 1999, and [Louwerse and Mitchell, 2003). The micro-level is concerned with relations within a turn or within a single utterance, such as rhetorical relations; the meso-level is about the relations within a sub-dialogue, e.g. units of grounding; and the macro-level is concerned with topic structure and plan-based analysis, topic shifts, and opening and closing a dialogue.

Speech production and editing

Speakers continuously monitor the utterance that is currently being produced, and when problems or mistakes are discovered, they stop the flow of the speech and signal to the addressee that there is trouble and that a repair follows (Clark and Krych, 2004). In natural conversation fluent speech is rare. Speakers make mistakes in verbal fluency, e.g. stuttering, or mispronouncing words, and may wish to reformulate part of an utterance or to start from the beginning of a phrase. Levelt (1989) mentions several reasons for repairs, such as lexical errors or flaws in formulation, syntactical or morphological errors, sound form errors, tongue slips, articulation errors, speaking style errors, and conceptual errors.

Allwood et al. (2005) introduced the term 'Own Communication Management (OCM)' to describe the activity of a speaker managing the planning and execution of his/her own communication, and argue that this is a basic function in dialogue. Similarly, Partner Communication

Management (PCM), introduced in (Bunt, 2006), is concerned with monitoring the partner's speech by a listener and providing assistance, e.g. by completing an utterance that the speaker is struggling to produce, or correcting (part of) a partner's utterance, believing that the partner is making a speaking error.

Timing

Planning takes time, as does the construction of communicative acts. Time management acts serve to allow a speaker to buy some time or to suspend his participation in the dialogue for a while. A dialogue participant who has the turn does not simply stop talking for some time for the necessary planning or construction work without indicating this to the addressee, because an unannounced silence creates uncertainty.

Clark (1996) notices that time delays can be signalled by modifying a syllable, word or phrase within a primary utterance using e.g. drawled syllables, non-reduced words, filled and silent pauses, or using other modalities (head movements, gaze direction, over-speech laughter, pointing). See also Bavelas and Chovil (2000), and Goodwin (1981) among others.

Concluding observations

To sum up, in the literature a range of aspects of communication is studied which involve communicative activities beyond those strictly related to performing the motivating task; notably actions concerned with auto- and allo-feedback, managing the use of time, taking turns, establishing contact, dealing with difficulties in utterance production, structuring the dialogue, and observing social aspects of interaction.

In the next section we investigate to what extent these aspects of communication are empirically observed in dialogue data.

3.5 Empirical observations from dialogue corpora

An analysis of the three corpora mentioned in Section 3.3 shows that the most frequently occurring category of dialogue acts is those that advance the *task* or *activity* that motivates the dialogue, see Table 3.1. Being multi-party, AMI meetings involve relatively much turn management, where participants perform dialogue acts to take the turn rather than just start speaking (more than half of all segments is preceded by certain turn-obtaining events (59%)); they interrupt each other (4.4%) and they speak simultaneously (20% of all segments partly overlap). The third largest category of functional segments in the AMI and DIAMOND corpora is *auto-feedback*. We observed that in AMI meetings one minute of conversation contains on average 9.4 positive auto-feedback segments. In OVIS dialogues a large portion of *allo-feedback* was observed, due to the fact that the OVIS system constantly checks the correctness of the output of its ASR module and the user reports back on the correctness of the system's understanding, thereby addressing the dimension of allo-feedback.

The distribution of the data across dimensions is one of the main distinguishing features of different dialogue types: multi- *vs* two-party, face-to-face *vs* remote, human-human *vs* human-machine, formal *vs* informal, and instructive *vs* information seeking *vs* meeting.

Table 3.1: Distribution of functional segments across dimensions for three dialogue corpora (in %).

	AMI	DIAMOND	OVIS
Task	31.8	45.1	48.8
Auto-Feedback	20.5	19.1	24.1
Allo-Feedback	0.7	3.8	39.2
Turn Management	50.2	19.9	19.3
Social Obligation Management	0.5	7.8	3.8
Discourse Structuring	2.8	2.3	3.3
Own Communication Management	10.3	0.7	3.4
Time Management	26.7	16.1	10.8
Partner Communication Management	0.3	0.3	0.5
Contact Management	0.1	2.8	12.3

Table 3.2: Inter-annotator agreement and tagging accuracy per dimension for the OVIS and DIAMOND corpora.

	Inter-annotator agreement (standard kappa)			Inter-annotator agreement (weighted kappa)		Tagging accuracy		
Dimensions	p_o	p_e	κ_{st}	p_e	κ_w	p_o	p_e	κ
Task	0.85	0.1	0.83	0.47	0.72	0.91	0.47	0.81
Auto-Feedback	0.91	0.1	0.9	0.24	0.88	0.94	0.24	0.92
Allo-Feedback	0.93	0.1	0.92	0.43	0.88	0.95	0.43	0.91
Turn Management	0.93	0.1	0.92	0.08	0.92	0.92	0.08	0.92
Time Management	0.99	0.1	0.99	0.11	0.99	0.99	0.11	0.90
Discourse Structuring	0.99	0.1	0.99	0.05	0.99	0.87	0.05	0.87
Contact Management	0.99	0.1	0.99	0.14	0.88	0.91	0.14	0.89
Own Communication M.	0.99	0.1	0.99	0.02	0.99	1.00	0.02	1.00
Partner Communication M.	0.99	0.1	0.99	0.002	0.99	1.00	0.002	1.00
Social Obligation M.	0.99	0.1	0.99	0.09	0.99	0.95	0.09	0.95

3.6 Dimension recognition

In order to assess the recognisability of dimensions we performed experiments where three expert annotators annotated DIAMOND and OVIS dialogues by assigning DIT^{++} tags. Table 3.2 presents inter-annotator agreement on dimensional tags for expert annotators expressed in terms of standard kappa (Cohen, 1960) and weighted kappa (see Geertzen and Bunt, 2006 and Chapter 4, Section 4.1), taking the class distribution into account (see Table 3.1), and tagging accuracy.[7] The table shows that there is near perfect agreement between annotators, and that accuracy is very high.

We used the rule induction algorithm Ripper (Cohen, 1995) which has been shown by Geertzen et al. (2007) to perform best on our data compared to statistical learners (e.g. Naive-Bayes classifiers) and memory-based learners (e.g. IB1). The features included in the data sets for dimension recognition relate to *dialogue history*: tags of the 10 (AMI and OVIS) or 4 (DIAMOND) previous turns; *prosody*: minimum, maximum, mean, and standard deviation of *pitch* (F0 in Hz), *energy* (RMS), *voicing* (fraction of locally unvoiced frames and number

[7] This is done by comparing the data produced by annotators with a gold standard (Geertzen et al., 2008).

Table 3.3: Dimension recognition scores in terms of accuracy (in %) comparing to baseline scores (BL) for each dimension and data set.

	DIAMOND data		AMI data		OVIS data	
Dimension	BL	Accuracy	BL	Accuracy	BL	Accuracy
Task	64.9	70.5	66.8	72.3	60.8	73.5
Auto-Feedback	71.1	85.1	77.9	89.7	66.1	75.9
Allo-Feedback	86.9	96.6	96.7	99.3	52.5	80.1
Turn Management	69.5	90.0	59.0	93.0	89.8	99.2
Time Management	65.6	82.2	69.7	99.4	95.5	99.4
Discourse Structuring	59.0	67.9	98.0	92.5	76.3	89.4
Contact Management	88.0	95.2	99.8	99.8	87.7	98.5
Own Communication Management	77.4	83.1	89.6	94.1	99.7	99.7
Partner Communication Management	45.4	62.6	99.7	99.7	99.8	99.8
Social Obligation Management	80.3	92.2	99.6	99.6	96.2	98.4

of voice breaks), and *duration*; and *word occurrence*: a bag-of-words vector[8] indicating the presence or absence of words in the segment. In total, 1,668 features are used for AMI data, 947 for DIAMOND data and 240 for OVIS data.[9]

Table 3.3 presents the resulting scores using the Ripper classifier obtained in 10-fold cross-validation experiments.[10] As the results show, the 10 dimensions defined in the DIT^{++} tag set are automatically recognizable with fairly good accuracy.

3.7 The independence of dimensions

The distinction of a dimension only makes sense if it can be separated clearly from the other dimensions that are considered. Therefore, in (Bunt, 2006) it was proposed as part of the definition of 'dimension' that it corresponds to an aspect of communication that a segment may address independently of other aspects that it might also address. This means that a segment may in principle be assigned any tag in a given dimension, regardless of whatever tags have been assigned to it in other dimensions. This is only *in principle*, though; empirically, there are restrictions on assigning tags in multiple dimensions. For example, accepting an offer cannot have a negative feedback function, because accepting presupposes that the speaker believes to have understood the offer; similarly, a farewell greeting closing a dialogue cannot have a feedback elicitation function. So the assignment of a communicative function in a certain dimension may impose restrictions on the possible tagging in another dimension. Such occasional restrictions on the co-assignment of tags in different dimensions correspond to empirical facts about communication, and do not affect the independence of the dimensions. Two dimensions are *not* independent if there are systematic relations between the tags in one dimension and those in the other, in particular if the tag in one dimension can be computed from that in the other.

We define the independence of dimensions as follows. First, we define the pairwise inde-

[8] With a size of 1,640 entries for AMI data, 923 for DIAMOND data and 219 for OVIS data.

[9] The features used in classification experiments will be discussed in Chapter 6, Section 6.1 in more detail.

[10] In order to reduce the effect of imbalances in the data, it is partitioned ten times. Each time a different 10% of the data is used as test set and the remaining 90% as training set. The procedure is repeated ten times so that in the end every instance has been used exactly once for testing and the scores are averaged. The cross-validation was stratified, i.e. the 10 folds contained approximately the same proportions of instances with relevant tags as in the entire dataset.

pendence of two dimensions:

(8) **Definition.** Two dimensions D_1 and D_2 are called *pairwise independent* iff:

1. a functional segment may have a D_2 function, regardless of whether it also has a D_1 function (and vice versa);
2. if a functional segment has both a D_1 function and a D_2 function, then the D_2 function is in general not determined by the D_1 function (and vice versa).

(9) **Definition.** The dimensions in a set D are independent iff every pair $< D_i, D_j > \in D$ is pairwise independent. Such a set of dimensions is also called *orthogonal*.

As mentioned in Section 3.3, we performed five dependency tests that assess (1) the relative frequency of communicative function co-occurrences across dimensions; (2) the extent of relatedness between dimensions measured with the phi coefficient; (3) dimension vector distances in multidimensional space; (4) for all dimensions whether there are functional segments addressing only that dimension; and (5) the relative frequency of pairs of dimension and previous dimension. The test results presented in this section are similar for all three studied corpora.

Function co-occurrences

Table 3.4 shows no dependences between dimensions, although some combinations of dimensions are relatively frequent, e.g. time and turn management acts often co-occur. A speaker who wants to win some time to gather his thoughts and uses Stalling acts mostly wants to continue in the sender role, and his stalling behaviour may be intended to signal that as well (i.e. to be interpreted as a Turn Keeping act). But stalling behaviour does not always have that function; especially an extensive amount of stallings accompanied by relatively long pauses may be intended to elicit support for completing an utterance. It can be also observed that functions which address the same dimension never co-occur, except for Auto- and Allo-Feedback where functions are not mutually exclusive but entail or implicate each other (see Chapter 5).

It is also interesting to have a look at co-occurrences of communicative functions taking implicated and entailed functions into account (the corpora were re-annotated for this purpose). As discussed in more detail in Chapter 5, in the case of an entailment relation, a functional segment has a communicative function, characterized by a set of preconditions which logically imply those of a dialogue act with the same semantic content and with the entailed communicative function. For instance, more specific functions entail less specific ones, such as Agreement and Disagreement entailing Inform, and Confirm and Disconfirm entailing Answer.

A communicative function in one dimension may also entail a function in another dimension. Such an entailment relation occurs for example between responsive acts in non-feedback dimensions on the one hand and feedback acts on the other.

Table 3.4: Co-occurrences of communicative functions across dimensions in the AMI corpus, expressed in relative frequency in %, implicated and entailed functions excluded and included. (Read as follows: percentage of segments having a communicative functions in the dimension corresponding to the column, which also has a function in the dimension corresponding to the row.)

	form	Task	Auto-F.	Allo-F.	Turn M.	Time M.	DS	Contact M.	OCM	PCM	SOM
Task	independent	0.0	1.1	0.0	2.2	0.1	19.6	0.0	3.8	0.0	0.0
	implied	49.8	47.9	24.9	97.5	2.4	31.5	0.4	69.6	0.1	0.7
Auto-F.	independent	0.7	0.0	0.0	11.0	0.6	1.9	11.1	0.8	0.0	0.0
	implied	38.9	100.0	0.0	88.7	11.4	11.2	20.2	11.7	65.0	8.7
Allo-F.	independent	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
	implied	24.9	0.0	100.0	94.8	35.7	2.1	1.2	7.9	0.7	0.3
Turn M.	independent	3.4	26.9	6.7	0.0	28.6	12.4	7.4	4.8	18.2	6.7
	implied	76.0	66.2	19.4	0.0	42.9	14.6	13.8	99.6	27.3	10.5
Time M.	independent	0.1	0.7	0.0	44.9	0.0	4.7	0.0	1.3	0.0	0.0
	implied	28.2	11.3	7.8	98.6	0.0	1.7	0.0	83.2	0.5	0.0
DS	independent	0.1	0.4	0.0	0.3	0.0	0.0	0.9	0.0	0.0	6.7
	implied	3.2	58.3	29.1	87.5	4.9	4.6	25.0	3.7	0.0	12.5
Contact M.	independent	1.7	0.3	0.0	3.6	0.5	3.7	0.0	0.0	0.0	1.3
	implied	2.4	97.1	1.6	98.8	0.5	2.4	0.0	0.3	0.0	3.7
OCM	independent	1.2	0.4	0.0	2.8	0.5	0.0	0.0	0.0	0.0	6.7
	implied	82.2	2.8	2.5	96.9	7.8	3.9	13.5	0.0	0.9	7.6
PCM	independent	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0
	implied	11.8	65.0	11.8	79.1	12.2	0.0	0.0	0.0	0.0	0.0
SOM	independent	0.0	0.0	0.0	0.2	0.0	0.0	2.7	0.3	0.0	0.0
	implied	0.7	80.0	10.0	90.0	0.0	30.0	3.9	2.0	0.0	0.0

A functional segment may also have multiple communicative functions due to the occurrence of conversational implicatures. For example, a shift to a relevant new discussion topic implicates positive feedback about the preceding discussion, while a shift to an irrelevant topic implicates negative feedback.

Co-occurrence scores are higher when entailed and implicated functions are taken into account. For example, questions, which mostly belong to the Task dimension, much of the time have an accompanying Turn Management function, either releasing the turn or assigning it to another dialogue participant, allowing the question to be answered. Similarly, for accepting a request the speaker needs to have the turn, so communicative functions like Accept Request will often be accompanied by functions like Turn Take or Turn Accept. Such cases contribute to the co-occurrence score between the Task and Turn Management dimensions.

Table 3.4 also shows that some dimensions do not occur in combination. We do not find combinations of Contact and Partner Communication Management, of Partner Communication Management and Discourse Structuring, or of Partner Communication Management and Social Obligation Management, for example. Close inspection of the definitions of the functions in these pairs of dimensions does not reveal any logical restrictions on the possible co-assignment of tags in these dimensions, hence these observation should be interpreted as empirical facts rather than as indications of dependences between the dimensions.

Inter-dimensional relatedness

Table 3.5 presents the extent to which dimensions are related when implicated and entailed functions are not taken into account (white cells) and when they are (gray cells), according to the calculated phi coefficient.

No strong positive (phi values from .7 to 1.0) or negative (-.7 to -1.0) relations are observed. There is a weak positive association (.6) between Turn and Time Management and between OCM and Turn Management (.4). Weak negative associations are observed between Task and Auto-feedback (-.5) when entailed and implicated functions are not considered; between Task and Contact Management (-.6); and between Auto- and Allo-feedback (-.6) when entailed and implicated functions are included in the analysis. A weak negative association means that a segment does not often have communicative functions in these two dimensions simultaneously. Some negative associations become positive when we take entailed and implicated functions into account because, as already noted, dialogue acts like answers, accepts and rejects imply positive feedback.

Dimension vector distances

For the third test we represented all annotated segments by vectors with 8 prosodic values (duration, min, max, mean, standard deviation in pitch, fraction voiced/unvoiced frames, voice breaks and intensity), 220 values for dialogue history and 1623 values for word tokens[11] occurring in the segment.

[11] Weights for tokens occurring in segments that have a communicative function in a particular dimension were computed by multiplying *token frequency* by the *inverse dimension frequency*.

Table 3.5: Extent of relatedness between dimensions for the AMI corpus expressed in the phi coefficient (implicated and entailed functions excluded (white cells) and included (gray cells)).

Dimension	Task	Auto-F.	Allo-F.	Turn M.	Time M.	DS	Contact M.	OCM	PCM	SOM
Task	-	.1	.3	.06	-.4	-.6	.03	-.03	-.1	.04
Auto-F.	-.5	-	-.6	.1	-.3	.2	-.02	.02	-.1	.04
Allo-F.	-.2	-.03	-	.09	-.1	-.2	0.3	-.01	-.02	-.01
Turn M.	-.03	-.04	.14	-	.6	.04	-.06	.02	.02	-.03
Time M.	-.4	-.06	.14	.6	-	-.1	-.02	.04	-.03	-.02
Contact M.	-.05	-.01	-.00	.00	-.01	-	.04	-.01	-.04	-.03
DS	-.2	-.02	-.01	-.01	-.02	-.00	-	-.01	-.01	.2
OCM	.01	-.05	.02	.4	-.03	-.01	-.00	-	-.03	-.01
PCM	-.1	-.01	.01	-.01	.01	-.00	-.01	-.01	-	-.003
SOM	-.1	-.01	-.01	-.02	-.02	-.00	.05	-.01	-.00	-

Table 3.6: Distances between dimensions.

Dimension	Task	Auto-F.	Allo-F.	Turn M.	Time M.	Contact M.	DS	OCM	PCM	SOM
Task	.000									
Auto-F.	82.911	.000								
Allo-F.	70.952	33.906	.000							
Turn M.	110.264	39.855	.267	.000						
Time M.	120.260	53.530	53.668	13.833	.000					
Contact M.	.979	132.211	148.877	171.665	.428	.000				
DS	87.027	136.307	141.244	174.537	187.414	57.944	.000			
OCM	110.561	42.106	44.225	3.597	11.951	173.970	176.066	.000		
PCM	92.694	31.326	32.236	19.186	28.982	161.210	159.736	19.543	.000	
SOM	33.101	72.339	65.130	104.994	116.671	101.148	80.440	105.827	90.655	.000

To simplify the distance measures between dimensions we constructed for each dimension a dummy dimension at the centre of the dimension cloud, which is the centroid C = $(c_1, c_2, ..., c_j)$, in which every c_j is the mean of all the values of j:

$$c_j = \frac{1}{N} \sum_{j=1}^{N} w_{i,j}$$

where w is the weight value for each feature. We then measured the distances between dimension vectors pair-wise using Euclidean distance:

$$euclid(\overrightarrow{d_j}, \overrightarrow{d_k}) = \sqrt{\sum_{j=1}^{n} (w_{ij} - w_{ik})^2}$$

Table 3.6 presents the results of calculating distance measures between centroid dimension vectors. There are no vectors which cross or overlap each other, although some dimension vectors are closer to each other in space, e.g. the Task dimension is relatively close to the Discourse Structuring dimension and Contact Management and Discourse Structuring because they share more or less the same vocabulary; Turn Management is close to Own Communication Management because they have similar prosodic properties, like duration and pitch; Turn and Time Management very often share the same vocabulary and some prosodic properties, like intensity and standard deviation in pitch.

Table 3.7: Overview of dimensions being addressed without any other dimension also being addressed in AMI, OVIS and DIAMOND data, expressed in relative frequency in %.

Dimension	Frequency (in %)		
	AMI	OVIS	DIAMOND
Task	28.8	37.9	29.9
Auto-Feedback	14.2	16.3	20.9
Allo-Feedback	0.7	4.1	6.8
Turn Management	7.4	0.9	8.5
Time Management	0.3	0.4	0.7
Contact Management	0.1	0.3	0.7
Discourse Structuring	1.9	1.8	2.7
Own Communication Management	0.5	0.8	2.7
Partner Communication Management	0.2	3.1	0.4
Social Obligation Management	0.3	6.4	0.7

Independent addressability

Concerning the very simple fourth test, Table 3.7 shows that each dimension may be addressed by a functional dialogue segment without any other dimension being addressed. This shows that each of the defined dimensions may be regarded as an autonomous aspect of communication.

Dimension sequencing

Finally, we investigated the occurrences of dimension tags given the tags of the previous segments taking 5 previous segments from the dialogue history. Table 3.8 shows that there is

no evident dependence in dimension relations across the dialogue history; there is no need for the speaker to address a particular aspect of communication as a response to the previous contributions.

Table 3.8: Overview of relative frequency (in %) of dimensions given the dimensions addressed by previous segments, observed in AMI data, per dimension, using the last 5 segments in the dialogue history.

Dimension	Task	Auto-F.	Allo-F.	Turn M.	Time M.	Contact M.	DS	OCM	PCM	SOM
Task	21.2	27.4	27.7	20.0	32.5	0.0	7.1	16.4	15.2	32.1
Auto-F.	15.0	24.4	25.0	21.4	15.4	27.8	12.3	7.5	22.7	12.8
Allo-F.	0.4	1.3	5.6	0.5	0.5	0.0	0.6	0.4	0.0	0.0
Turn M.	14.3	4.7	0.0	6.5	5.2	0.0	6.5	2.2	7.6	6.4
Time M.	22.2	16.3	16.7	23.5	15.0	0.0	35.5	47.1	37.9	19.2
Contact M.	0.0	0.1	0.0	0.2	0.0	27.8	0.0	0.0	0.0	0.0
DS	2.0	2.0	0.0	0.5	0.5	27.8	5.2	0.4	0.0	0.0
OCM	7.7	6.3	5.6	7.7	11.2	0.0	0.0	7.0	0.0	0.0
PCM	0.4	0.4	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0
SOM	0.1	0.3	0.0	1.2	0.1	0.0	0.6	0.0	0.0	6.4

Some patterns can certainly be observed. For example, retractions and self-corrections often follow hesitations; this is because the speaker, while monitoring his own speech and noticing that the segment of part of what he just produced needs revision, often needs a while to reconstruct this segment. It was also observed that the co-occurrence scores for Turn Management, Task and Auto-Feedback with other dimensions are relatively high. This means that Task functional segments are frequently preceded or followed by Turn Management or Auto-Feedback segments or segments that have functions in these two dimensions simultaneously. For instance, a frequent pattern for constructing a turn is first performing a turn-initial act (e.g. Turn Take, Accept or Grab) combined with or followed by an Auto-Feedback act and one or more segments in another dimension, and closing up the turn with a turn-final act.

3.8 Dimension-related concepts in existing dialogue act annotation schemes

The comparison of existing annotation schemes was performed by inspecting the definitions proposed in the relevant manuals as well as examples from annotation guidelines and annotated corpus data. Some schemes do cover all facets of dialogue interaction considered in previous sections and have an almost one-to-one correspondence with theoretical distinctions. Others, by contrast, discard some aspects for practical reasons, e.g. in order to simplify the annotation task or to fulfil domain constraints that do not require elaborate and domain-independent dialogue modelling.

Task and Task Management

Multidimensional dialogue act taxonomies, such as **DAMSL**, **Coconut**, **MRDA**, **DIT**$^{++}$ and **LIRICS**, define a *Task* dimension for those dialogue acts that advance the task (or 'activity') that motivates the dialogue. **DAMSL** has two separate dimensions for this aspect, *Task*

and *Task Management* ('about task' in **MRDA** and **SWBD-DAMSL**). The latter explicitly addresses the way in which the task is performed and interpreted. The **MRDA** category 'about-task' covers similar information applied to meetings, and is defined as 'reference to meeting agendas or direction of meeting conversation'.

Table 3.9: Positive and negative **Auto-feedback** functions in existing dialogue act taxonomies.

LIRICS	**Positive Auto-Feedback**				
DIT^{++}	Positive attention	Positive perception	Positive interpretation	Positive evaluation	Positive execution
DAMSL	Signal understanding		Acknowledgment		
SWBD-DAMSL	Signal understanding		Acknowledgment Repeat-rephrase		Summarize-reformulate
MRDA	Signal understanding		Acknowledgment Appreciation	Assessment	
Coconut	Signal understanding		Acknowledgment Repeat-rephrase		
AMI	Comment-about-understanding POS			Assess	Inform POS
HCRC MapTask			Acknowledgment		
Verbmobil	Backchannel			Acknowledge	Positive feedback
SLSA	Pos.contact	Pos.perception	Pos.understanding	Pos. acceptance/attitude	
TRAINS	Acknowledgment			Pos.evaluation	
SPAAC	Echo		Acknowledge	Appreciate	
MALTUS	Pos. attention	Repeat-rephrase		Appreciation	
Chiba	Follow up: pos. understand			Pos. response	
Alparon			Acknowledgment		
C-Star			Acknowledgment		

LIRICS	**Negative Auto-Feedback**				
DIT^{++}	Negative attention	Negative perception	Negative interpretation	Negative evaluation	Negative execution
(SWBD-) DAMSL	Signal-non-understanding				
MRDA	Signal-non-understanding		Understanding Check		
Coconut	Signal-non-understanding		Clarification Check		
AMI	Comment-about-understanding NEG				Inform NEG
HCRC MapTask	Check				
Verbmobil	Request clarify				Neg.feedback
SLSA	Neg.contact	Neg.perception	Neg. understanding	Neg. attitude	
TRAINS			Neg. evaluation		
SPAAC	Pardon				
MALTUS	Neg.attention				
Chiba	Follow up: understand			Neg. response	
Alparon					
C-Star					

Feedback

Feedback is reflected in all existing dialogue act taxonomies except for **Linlin** (Dahlbaeck and Jonsson, 1998) and **Primula** (Popescu-Belis, 2004). In the majority of schemes various levels of feedback are defined, ranging from merely hearing what was said to identifying the speaker's intention.

For example, in **SLSA** (Nivre et al., 1998) a distinction is made between *giving* and *eliciting* feedback at the levels of *contact*, *perception* and *understanding*, which are comparable with the levels defined in **DIT**$^{++}$ (Bunt, 1994 and Bunt, 2006) as *attention*, *perception* and *interpretation*. The **AMI** scheme (AMI Consortium, 2005b) defines the *assess* function to express evaluative feedback (*assessment/appreciation* in **SWBD-DAMSL** and **MRDA**, *positive evaluation* in **TRAINS**, *acknowledge* in **Verbmobil**); for expressions of auto-feedback concerning perception and interpretation, **AMI** has *comment-about-understanding*; **DAMSL**, **MRDA**, **Coconut**, and **HCRC Maptask** have *acknowledgment*. Table 3.9 gives an overview of the communicative functions defined for Auto-Feedback in the various schemes.

Dialogue participants monitor each other for their understanding and evaluation, and when necessary correct each other or elicit feedback (Clark and Krych, 2004). **LIRICS**, **DIT**$^{++}$, **SLSA** and some other schemes make a distinction between auto-feedback, which is about the speaker's processing of the previous discourse, and *allo-feedback*, which is about the addressee's processing. **SWBD-DAMSL** and **MRDA** define *backchannels in question form* for utterances like *'right?'*. **MRDA** has also *'follow-me'* questions where the speaker wants to verify that he is being understood, e.g. *'Do you know what I mean?'*. The **AMI** scheme includes several feedback elicitation functions: *elicit inform*, *elicit assessment*, and *elicit comment-about-understanding*. Feedback elicitation is defined in 12 of the 18 analysed annotation schemes.

Table 3.10: **Turn Management** functions in existing dialogue act taxonomies.

LIRICS	Take	Grab	Accept	Keep	Assign	Release
DIT^{++}	Take	Grab	Accept	Keep	Assign	Release
DAMSL& Coconut				Turn maintain		
SWBD-DAMSL			Hold before answers	Turn maintain	Turn exit	
MRDA	Regain turn	Grabber	Hold before answers	Holder		
SLSA	Turn take	Interruption	Turn opening	Turn holding	Turn closing	
TRAINS	Take			Keep	Assign	Release
SPAAC				Hold		
MALTUS& PRIMULA		Turn grabber		Turn holder		Back-channel
Chiba				Hold		

Taking Turns

Most dialogue act annotation schemes include communicative functions for dealing with turn management (see Table 3.10). Turn-initial acts are concerned with accepting, taking, or grabbing the speaker role; turn-final acts with releasing the speaker role or assigning it to someone else.

A speaker may also want to stay in the speaker role. In this case, no reallocation of the speaker role occurs. The activities that are performed in order to achieve this constitute a *turn keeping* act (also called turn *maintaining* or *holding*).

Social obligations and politeness

As Table 3.11 shows, all dialogue act annotation schemes include functions for social obligations and politeness, except for **Chiba** (Ichikawa, 1998) and **HRCR Maptask** (Carletta et al., 1996). **AMI** and **Verbmobil** have unspecified social obligation functions; for example, *politeness_formulae* in **Verbmobil** includes any social acts of politeness like greeting, apologizing, thanking, but also good-natured jokes, positive comments or compliments.

Table 3.11: **Social Obligation Management** functions in existing dialogue act taxonomies.

DIT^{++}	Greeting/ return greeting	Self-introduction/ return self-introduction	Goodbye/ return goodbye	Apology/ accept apology	Thanking/ accept thanking
LIRICS	Greeting/ return greeting	Self-introduction/ return self-introduction	Goodbye/ return goodbye	Apology/ accept apology	Thanking/ accept thanking
DAMSL	Greeting		Goodbye		
SWBD-DAMSL	Greeting			Apology/ downplayer	Thanking/ downplayer
MRDA				Downplayer/ sympathy	Thanking
Coconut	Greeting		Goodbye		
AMI	Be-positive/be-negative				
Verbmobil	Greet	Introduce	Bye	Politeness formulae	Thank
SLSA	Greet				
TRAINS	Greet				
MALTUS	Politeness				
Primula	Politeness; face-threatening/face-saving				
Alparon	Greet		Bye		
C-Star	Greeting	Self-introduction		Apologize	Thanking

Discourse and topic structure

Except for **AMI**, **TRAINS** and **Alparon** all schemes include communicative functions for explicitly structuring the discourse. In **AMI** separate taxonomies have been designed (see Xu et al., 2005 and Rienks and Verbree, 2005) to analyse topical and argumentative structures in meetings.

Opening and *closing* are the most frequently defined functions in this dimension. Some schemes include unspecified functions for topic management (**Coconut, Linlin** and **SPAAC**); others have more specific functions such as *topic change/shift* (**DIT^{++}** and **MRDA**), or *ready* (**HCRC Maptask**); *topic introduction/opening* (**SLSA**, **C-Star**).

Monitoring one's own and the addressee's speech

As Table 3.13 shows, 10 of the 18 analysed schemes include functions for monitoring and editing one's own speech (*own communication management*). **DAMSL** and **Coconut** deal with

Table 3.12: **Discourse Structuring** functions in existing dialogue act taxonomies.

DIT^{++}	Opening	Pre-closing	Topic introduction	Topic shift	Topic shift announcement
LIRICS	Interaction structuring				
DAMSL	Opening	Closing			
SWBD-DAMSL	Opening	Closing			
MRDA			Topic change		
Coconut	Opening	Closing	Topic		
AMI	Argument structure and topic segmentation schemes				
HCRC MapTask			Ready (for topic shifts)		
Verbmobil		Task close	Task initiate	Digress	
LinLin	Opening	Closing	Topic layer		
SLSA	Opening	Closing	Opening	Continuation	
SPAAC			Initiate: release issue	Topic	
MALTUS				Topic change	
Primula	Opening	Closing	Topic opening	Topic closing/change	
Chiba	Opening	Closing	Topic break		
C-Star		Closing	Introduce topic		

this phenomenon in their Communication Management dimension, without defining specific communicative functions. **Coconut** includes the *correct assumption* function for both semantic partner- and self-corrections.

Partner communication management is concerned with monitoring the partner's speech, providing assistance by completing an utterance that the partner is struggling to produce (*completion*) or correcting (part of) the partner's utterance. **MALTUS** (Popescu-Belis, 2004) defines the *restated info with correction* function, leaving unspecified whether speaker or partner is corrected.

Table 3.13: **Own** and **Partner Communication Management** functions in existing dialogue act taxonomies.

	Own Communication Management			Partner Communication Management	
DIT^{++}	Error signalling	Retraction	Self-correction	Correct-misspeaking	Completion
LIRICS	Error signalling	Self-correction		Correct-misspeaking	Completion
DAMSL		Speech repair		Correct-misspeaking	Completion
SWBD-DAMSL		Speech repair		Correct-misspeaking	Completion
MRDA		Speech repair		Correct-misspeaking	Collaborative completion
Coconut		Correct assumption; Speech repair		Correct-misspeaking	Completion
SLSA	Change				
TRAINS		Repair			
SPAAC		Correct-self		Correct	Complete
MALTUS		Restated info with repetition/correction		Restated info with correction	

Time

Most (viz. 12) of the analysed schemes define dialogue acts that address the management of time in dialogue. *Stalling* is the function of utterances where the speaker signals that he needs

a little time to produce a contribution. **AMI** defines *stallings* as special cases; it is argued that these utterances are not really a dialogue act, since the speaker does not necessarily convey an intention in these segments. **SLSA** has *choice* as a mechanism enabling the speaker to gain time for processes having to do with the continuation of the interaction (involving hesitation, memory search, planning, and keeping the floor), but these are thought to address the OCM dimension. In **TRAINS** this function is covered by the *turn-maintaining* tag.

Three tendencies may be observed here: (1) Time Management defined as a separate class on its own; (2) time-related acts defined but considered as unintentional acts; and (3) time management considered as part of the functions for Turn Management or Own Communication Management.

Contact and attention

In 6 of the 18 studied dialogue act schemes tags are defined for addressing the monitoring of contact and attention. **DAMSL**, **SWBD-DAMSL** and **Coconut** have *communication channel* establishment in the Communication Management dimension, for utterances like *'Are you there?'*. **Verbmobil** defines a *refer-to-settings* tag which addresses the settings of interaction, e.g. noise in the room, or the output quality of the computer used in the interaction. **HRCR Maptask** has *align* for checks of the attention or agreement of the partner, or his/her readiness for the next move (the second part of the definition is particularly relevant here).

Table 3.14: **Time** and **Contact Management** functions in existing dialogue act taxonomies.

	Time Management		Contact Management	
DIT^{++}	Stalling	Pausing	Contact check	Contact indication
LIRICS	Stalling	Pausing	Contact check	Contact indication
DAMSL	Communication management: delay		Communication channel	
SWBD-DAMSL	Stalling; delay; Hold before answers		Communication channel	
MRDA	Hold before answers			
Coconut	delay		Communication channel	
AMI	Stall			
Verbmobil	Deliberate		Refer-to-settings	
SLSA	Choice			
TRAINS	Keep			
SPAAC	Hold			
Alparon		Pause		
C-Star		Please wait		

Summary

To summarize, the following aspects of communication are reflected in the analysed dialogue act annotation schemes as follows: **Task**: present in all schemes except in SLSA; **Auto-Feedback**: present in 16 schemes; **Allo-Feedback** (elicitation): in 12 schemes; **Turn Management**: in 12 schemes; **Discourse Structuring**: in 16 schemes; **Social Obligation Management**: in 16 schemes; **Own Communication Management**: in 10 schemes; **Partner Communication Management**: in 8 schemes; **Time Management**: in 12 schemes; **Contact Management**: in 6 schemes.

3.9 Summary

In this chapter we have discussed the notion of dimension as an aspect of communication which an utterance can address in a dialogue context. Five criteria were defined for including a dimension in an annotation scheme: (1) theoretically and (2) empirically motivated; (3) recognized by human annotators and automatically; (4) reflected in existing annotation schemes; and (5) independently addressable. Table 3.15 gives an overview of the results of our investigations with respect to these criteria.

Table 3.15: Summary of survey and testing results in identifying a proper set of dimensions.

	Theoretical validation	Empirical justification	Recognisability	Compatibility	Independent addressability
Task	+	+	+	+	+
Task Management	+	+	-	+	na
Auto-Feedback	+	+	+	+	+
Allo-Feedback	+	+	+	+	+
Turn Management	+	+	+	+	+
Social Obligation M.	+	+	+	+	+
Own Communication M.	+	+	+	+	+
Discourse Structuring	+	+	+	+	+
Partner Communication M.	+	+	+	+	+
Time Management	+	+	+	+	+
Contact Management	+	+	+	+	+

The analysis shows that ten dimensions can be considered as good candidates to be used in an annotation scheme, namely **Task**, **Auto-Feedback**, **Allo-Feedback**, **Turn Management**, **Social Obligations Management**, **Own Communication Management**, **Discourse Structuring**, **Partner Communication Management**, **Time Management** and **Contact Management**. They have been studied extensively, from both theoretical and practical points of view; they are observed in actual dialogues; they are reliably annotated and successfully classified automatically; they are defined in most existing annotation schemes; and they address a certain aspect of communication independently of others. The results of this study have been the basis for choosing the nine dimensions of the ISO dialogue act annotation standard.

Distinguishing these dimensions does not mean that a dialogue act annotation effort should necessarily use all ten dimensions. An annotator or analyst who is especially interested in certain aspects of communication can choose to use only the corresponding dimensions of the scheme. Depending on the annotation task, some dimensions can be left out. For example, **Contact Management** is an important aspect in some types of dialogue, such as telephone conversations or tele-conferences (as in the OVIS corpus), but may be rather insignificant in face-to-face interaction. On the other hand, an annotator or analyst who is interested in an analysis that is not covered by this set can add a dimension, provided that it satisfies the condition of orthogonality with respect to the other dimensions.

A comprehensive multidimensional annotation scheme like DIT^{++} or as defined in ISO standard 24617-2 cannot be expected to be ideal for every kind of dialogue analysis, for every task domain, for every kind of dialogue, and for every annotation purpose. The general principles underlying the design of the scheme and the DiAML annotation language should however be useful for accommodating extensions, modifications, or restrictions of the scheme and the annotation language, as the need arises for particular applications. The main principles for annotation scheme extension or restrictions with respect to dimensions can be formulated

as follows:

- Any dimension and the corresponding set of dimension-specific communicative functions may be freely *left out*.
- For specific purposes or domains, new dimensions may be *added*:
  - the requirement of theoretical justification does not need to be observed, since the purpose may be to investigate dialogue phenomena which have not been studied yet;
  - the requirement that the dimension should occur in a significant number of annotation schemes may be dropped when adding a dimension for a specific purpose or a particular domain;
  - the resulting set of dimensions should remain orthogonal as much as possible.

Chapter 4

Dialogue act annotation

The main aim of this chapter is to go into details of dialogue act annotation practices. A number of fundamental issues are discussed, as well as practical aspects of transcription and segmentation. Existing approaches to dialogue act annotation are discussed, with their commonalities and differences, as well as their advantages and shortcomings. In particular, one-dimensional and multidimensional approaches are contrasted. Some improvements are proposed in order to deal with phenomena like certainty, conditionality and sentiment to better fit multimodal data. We describe in detail annotation work performed using the DIT^{++} tag set, discussing dialogue corpus data and issues in dialogue segmentation.

Introduction

Recent years have witnessed a growing interest in annotating linguistic data at the semantic level including annotation of dialogue corpus data with dialogue act information. For this purpose a variety of dialogue act annotation schemes have been developed. Many of these schemes were designed for a particular purpose or a particular domain.

The MATE project (Klein and Soria, 1998) was one of the first to address issues in standardisation of dialogue act annotation concepts. It aimed to contribute to the development of standards for annotating resources and to provide methods for improving the efficiency of knowledge acquisition and extraction processes. A set of tools was designed for mapping, extraction, visualization and evaluation of annotated dialogue data within the compared approaches has been designed.

Larsson (1998) made a comparative analysis of three schemes: DAMSL, LinLin and HCRC MapTask. He noticed that these schemes differ with respect to the range of phenomena they cover, the division of phenomena into levels or layers, the division of layers into categories, and their domain and genre dependency. The schemes also differ in the types

This chapter is largely based on Petukhova and Bunt (2007); Petukhova and Bunt (2010b); and Petukhova et al. (2011). These conference papers are mostly written by me with ample support from my co-authors. The conclusions of this chapter contain some discussions originating from the ISO 24617-2 project, in close cooperation with other editorial group members. My part in this project was substantial and I was involved in all parts and aspects of the project.

of definition they provide; some schemes have surface-based definitions like HCRC Map-Task (e.g. Query-yn or Reply-y), others provide intention-based definitions. Some schemes make use of instructional definitions, e.g. in DAMSL we see definitions given informally as instructions for annotators. Larsson calls these differences 'dimensions of variation' which should be taken into account when designing a dialogue act annotation scheme, and lists the following ones: coverage of phenomena, division into layers, division into categories and subcategories, segmentation principles, relational tags, multifunctionality of utterances, multi-agent acts, discontinuous utterances, domain dependency, dialogue genre dependency, theory dependency, and definition types. Larsson (1997) proposed a scheme with two levels of hierarchy. The top level contains core speech acts (initiatives and responses), feedback/grounding moves (eliciting and providing feedback), turn-taking moves and 'conventional moves'. The lower level has clusters populated with communicative functions constructed by combining the DAMSL, HCRC, LINLIN, TRAINS and GBG-IM coding schemes, e.g. the Agreement (DAMSL) cluster, which belongs to the responsive core speech acts, contains the following communicative functions: Accept (DAMSL/TRAINS), +Accept-content (GBG-IM), Accept-part (DAMSL), Maybe (DAMSL), Reject (DAMSL/TRAINS), +Reject-content (GBG-IM), Reject-part (DAMSL), Hold (DAMSL). This use of hierarchies and clusters of semantically related functions makes the comparison more feasible.

Soria and Pirrelli (2003) describe an approach to dialogue act scheme mapping, making use of a "meta-scheme". A meta-scheme is a framework for comparing schemes along orthogonal dimensions of linguistic or contextual analysis which have a bearing on the definition of dialogue acts. They claim that comparing annotation schemes via a meta-scheme enables a judgment about the similarity of tag sets. This approach was discussed in the previous chapter 3.

Popescu-Belis (2004) made an empirical comparison of the DAMSL, SWBD-DAMSL, and ICSI-MR dialogue act tag sets. He noticed that the number of possible DAMSL tags is very large (about 4 million possible combinations), because DAMSL does not have many mutually exclusive functions. SWBD-DAMSL brought important changes to the DAMSL set by collapsing layers and dimensions, and introduced exclusivity constraints resulting in a set of 220 unique tags. Popescu-Belis investigated the tag combinations which occur in annotated data, eliminated infrequently used combinations of tags, and reduced the SWBD-DAMSL set to 42 mutually exclusive tags. The ICSI-MR dialogue act tag set, which is itself a multidimensional version of the SWBD tag set, was analysed and converted into the MALTUS tag set (Multidimensional Abstract Layered Tagset for Utterances), using the most frequent ICSI-MR tag combinations and maintaining high tagging accuracy (0.061% error rate). 26 labels were finally proposed in the MALTUS scheme.

Another empirical approach to dialogue act annotation scheme comparison was proposed by Petukhova (2005). Three schemes were compared, two multidimensional (DAMSL and DIT^{++}) and one single-dimensional (AMI). The same dialogues were annotated according to each of the schemes, and the correspondences between assigned tags were analysed. Once the annotated data is available, systematic differences and correspondences between schemes, and their strengths and weaknesses become apparent. As a method for schemes comparison this approach has practical disadvantages, however. First of all, a large corpus is needed to make sure that all the labels defined in different schemes are present in annotations. Secondly, multiple annotators should be trained to apply various annotation schemes reliably.

In the LIRICS[1] project, methodological factors were studied which should be taken into

[1] Linguistic InfRastructure for Interoperable ResourCes and Systems (http://lirics.loria.fr)

consideration when isolating concepts for semantic annotation. A set of concepts for dialogue act annotation was defined in the form of ISO data categories for dialogue act annotation. This set has been tested for its usability and coverage in the manual annotation of dialogues in English, Dutch and Italian (see Bunt et al., 2007b).

In a collaborative effort, several research groups have embarked on the definition of a common framework for the design of embodied conversational agents (ECAs). The AAMAS workshops 'Towards a Standard Markup Language for Embodied Dialogue Acts' in 2008 and 2009 have led to the first steps in the definition of a standard Functional Markup Language (FML) for ECAs.[2] A major concern is that of defining the types of dialogue act to be performed by ECA systems. It was concluded that existing dialogue act taxonomies can be used for the development of FML, such as Conversation Acts (Allen et al., 1994), the DAMSL annotation scheme (Allen and Core, 1997) and the DIT^{++} taxonomy (Bunt, 2009a). The main difference between existing schemes and what is required for FML, is that for most of these schemes certain extensions are required, for example for dealing facial expressions and gestures of an ECA to express emotions and attitudes for which most existing schemes have no provisions.

This chapter is organised as follows. First, we discuss approaches to dialogue act annotation reflected in the dialogue act taxonomies, mentioned in Chapter 3 where dimension-related concepts distinguished in these tag sets were described. Section 4.2 discusses the segmentation of dialogue into units that are relevant for dialogue analysis, and presents an approach to segmentation in multiple dimensions. Section 4.3 studies relations between dialogue units. Section 4.4 discusses communicative functions qualification. Section 4.5 goes into details of the annotation work that was performed as part of this thesis describing the data and introducing the semantic framework of Dynamic Interpretation Theory (DIT), in particular the DIT^{++} taxonomy (Section 4.5.2). Section 4.6 summarises the basic concepts used in dialogue act annotation, in line with the ISO metamodel for dialogue act annotation (ISO DIS 24617-2:2010).

4.1 Approaches to dialogue act annotation

Dialogue act annotation schemes can be divided into one- and multidimensional ones. One-dimensional schemes propose tag sets which are as a rule kept fairly simple, are mostly used for coding dialogue utterances with only one tag. Some one-dimensional schemes propose a list of mutually exclusive communicative functions as a tag set, such as Alparon (Van Vark et al., 1996), precluding the assignment of multiple tags to an utterance. Other schemes cluster semantically related functions into groups, such as Verbmobil (Alexandersson et al., 1998) and AMI (AMI-Consortium, 2005b). The introduction of clusters of tags improves the transparency of the tag set and makes the coverage of the tag set clearer, since each cluster typically corresponds to a certain class of dialogue phenomena.

A grouping based on semantic similarity is, however, not sufficient to support a satisfactory account of multifunctionality as we showed in Chapter 3. A multidimensional annotation scheme provides a systematic way to capture the multifunctionality of dialogue utterances and to support the decision-making process for annotators in dealing with multifunctional utterances. For instance, the DAMSL guidelines include a procedure for annotating multifunctional utterances which instructs the annotator to consider potential utterance functions in four layers: Communicative Status, Information Level, Forward-Looking Function and Backward-Looking Function. This results in more efficient and consistent assignment of multiple tags.

[2] Detailed information can be found at http://hmi.ewi.utwente.nl/conference/EDAML

It has been noted (e.g. Klein, 1999; Larsson, 1998) that one-dimensional annotation schemes have serious disadvantages. Allen and Core (1997) and Allwood (2000a) note that a set of mutually exclusive categories cannot account for the fact that utterances may perform multiple actions simultaneously, and in this respect they also criticise traditional speech act theory.

Existing one- and multidimensional schemes differ not only in their precise sets of tags, but more importantly with respect to (1) definition of the related concepts; (2) level of granularity of the tag set; and (3) coverage of relevant dialogue phenomena.

Chapter 3 discussed the 18 most widely-used dialogue act annotation schemes and provided an overview and mapping of dimension-specific communicative functions. The comparison of general-purpose communicative functions defined in these annotation schemes is presented in Bunt et al. (2010). For the majority of taxonomies, definitions of the dialogue act types are intention-based, e.g. DAMSL, DAMSL-based schemes, DIT^{++}, AMI and Verbmobil. Definitions, however, are informal and descriptive, and are often given in the form of instructions for the annotator. For these schemes, DIT^{++} is the only one that provides precise definitions and fine-grained distinctions between communicative functions. Other schemes provide form-based (or feature-based) definitions, e.g. SPAAC, some definitions in the MRDA and HCRC MapTask coding schemes.

Multidimensional schemes as a rule have better coverage of dialogue phenomena than one-dimensional schemes. For instance, in AMI no communicative functions are defined for aspects of interaction management such as time, topic, contact, own and partner communication management. Corpus analyses in (Petukhova, 2005) showed that these functions need to be included because this information is (1) a significant part of natural human conversation in general, and meetings in particular; and (2) important for understanding the functions of nonverbal acts (see Chapter 6). The DAMSL scheme does not define these classes of communicative functions either but can label the corresponding behaviour as being concerned with Communication Management.

Some schemes define tags based on fine-grained theoretical and empirical distinctions. For example, DIT^{++} defines 7 information-providing communicative functions (DAMSL has 11) based on differences in the speaker's motivation for providing the information, and additional beliefs about what the addressee knows. The AMI scheme defines one communicative function INFORM which can be combined with 4 relation tags: POSitive, NEGative, PARTial and UNCertain. This allows to annotate several types of answers, e.g. positive or negative answer, or positive uncertain answer, etc., but does not allow to differentiate between, for example, a confirm, an agreement and a positive propositional answer, or between those and accept request, accept suggestion and accept offer, which are not concerned with the exchange of information in propositional form, but address the performance of events.

One of the advantages of multidimensional annotation schemes is that they are more easily adapted to various purposes and task domains. For instance, DIT^{++} and DAMSL, initially designed for two-agent task-oriented dialogues, perfectly fit the AMI meeting data. Multidimensional schemes can be used for studying specific phenomena, such as the turn-taking behaviour in conversations, the roles of participants and their dominance relations, or the efficiency of a discussion. Only the relevant dimensions need to be considered and the others can be simply left out, without compromising the consistency of annotations.

Conversely, new dimensions and new elements within existing dimensions can be added in multidimensional taxonomies without affecting the rest of the scheme, provided that certain general design principles are followed, see Chapter 3. A multidimensional scheme may also

have open classes, allowing suitable additions of those communicative functions which are specific for a certain application, task domain, and modalities. For example, DAMSL defines open classes like Other Level to be extended with dimensions other than Task, Task Management or Communication Management, and Other Forward-Looking functions to be extended with functions that are not defined in DAMSL.

A multidimensional scheme is in principle straightforward to convert into a tag set with mutually exclusive complex tags. For instance, Popescu-Belis (2005) analysed 113,560 dialogue utterances (the ISCI-MRDA corpus) according to the multidimensional MRDA-annotation scheme and observed that about 760 of the approximately 7 million theoretically possible combinations occur in the corpus. The scheme can therefore be converted into an unstructured list of 760 complex tags. Further analysis of the frequencies of these labels and the dependencies between tags can reduce the space significantly, resulting in a theoretically and empirically well-motivated one-dimensional tag set. Note that an annotation scheme constructed in this way is not one-dimensional. The complexity of the tags is just another way of encoding multiple aspects of interaction.

Manual annotation is time consuming, and it is generally thought that tagging dialogue utterances according to a multidimensional scheme costs more time than when using a one-dimensional scheme. The analyses reported in (Petukhova, 2005) showed that the ratio of annotation time to real dialogue time ratio was approximately 25:1 when coding with the AMI scheme[3], and approximately 19:1 when coding with DIT^{++} or DAMSL. This can be explained by the fact that a one-dimensional annotation scheme poses quite a challenge for annotators, because it is often hard to judge what phenomena have been merged in a single tag.

An argument that is sometimes used against multidimensional schemes is that annotation using such schemes is not reliable, in terms of inter-annotator agreement. For measuring inter-annotator agreement, the standard kappa statistics or Krippendorff's α for multiple annotators are often used (see Cohen, 1960; Krippendorff, 1980 and Carletta, 1996). Reidsma (2008) reports the inter-annotator agreement when using AMI tags in terms of Krippendorff's α; α values range between 0.55 and 0.61. This is interpreted as indicating that the annotators reached a moderate agreement. For DAMSL, inter-annotator experiments were performed by Stent (2000) using 8 dialogues of the MONROE corpus, counting 2897 utterances in total, processed by two annotators for 13 DAMSL dialogue act clusters. The inter-annotator agreement was measured using standard κ, results are given in Table 4.1.

For the majority of main DAMSL dimensions near perfect agreement was achieved (influence - on-listener, influence-on-speaker, info-request, agreement, answer, statement and understanding).

For DIT^{++}, we reported the scores for dimension labelling in Chapter 3, Section 3.6, Table 3.2. Geertzen & Bunt (2006) measured the inter-annotator agreement for communicative functions for the 10 DIT^{++} dimensions. They noticed, when considering inter-annotator agreement for the use of hierarchically structured tag set, that the standard kappa statistic is not an appropriate measure, see also Lesch et al. (2005) where a hierarchy-based distance metric is proposed. Instead, a weighted kappa statistic was adopted which can take into account a probability distribution typical for each annotator, generalize it to the case for multiple observers by taking the average over the scores of annotator pairs, and which uses a distance metric taking the structure of the taxonomy of tags into account. Geertzen & Bunt (2006) proposed using weights in Cohens κ_{wt}. Weights are determined by the distance between tags in a hierarchy. A

[3]Annotation time for the AMI scheme has been measured by annotators at the University of Twente and reported in an internal report which is not publicly available.

Table 4.1: Inter-annotator agreement of DAMSL dialogue acts on the MONROE corpus. (adopted from Stent, 2000)

Category	p_o	p_e	κ
influence-on-listener	0.97	0.77	0.88
influence-on-speaker	0.95	0.73	0.83
info-request	0.98	0.81	0.90
agreement	0.96	0.60	0.89
answer	0.98	0.85	0.86
conventional	1.00	0.99	0.88
exclamation	0.99	0.98	0.70
info-level	0.88	0.70	0.59
other-forward-looking	0.99	0.90	0.88
performative	1.00	0.99	0.45
response-to	0.90	0.40	0.83
statement	0.93	0.41	0.88
understanding	0.96	0.56	0.91

coefficient that is called *taxonomically weighted kappa* is proposed and denoted by κ_{wt}:

$$\kappa_{wt} = 1 - \frac{\sum(1-\delta(i,j)).p_{oij}}{\sum(1-\delta(i,j)).p_{eij}}$$

where δ is a distance metric that measures disagreement and is a real number normalized in the range between 0 (not related functions) and 1 (identical functions). Table 4.2 shows the results, and compares standard and taxonomically weighted kappa scores.

Table 4.2: Standard and taxonomically weighted kappa statistics per dimension drawn from the set of all annotation pairs from the DIAMOND corpus. (from Geertzen & Bunt, 2006)

Dimension	standard			weighted		
	p_o	p_e	κ	p_o	p_e	κ
Task	0.52	0.09	0.47	0.76	0.17	0.71
Auto-Feedback	0.32	0.14	0.21	0.87	0.69	0.57
Allo-Feedback	0.53	0.19	0.42	0.79	0.50	0.58
Turn Management	0.90	0.42	0.82	0.90	0.42	0.82
Time Management	0.91	0.79	0.58	0.91	0.79	0.58
Contact Management	1.00	0.53	1.00	1.00	0.53	1.00
Own Communication Management	1.00	0.50	1.00	1.00	0.95	1.00
Partner Communication Management	1.00	1.00	nav	1.00	1.00	nav
Dialogue structuring	0.87	0.48	0.74	0.87	0.48	0.74
Social Obligation Management	1.00	0.19	1.00	1.00	0.19	1.00

From the agreement scores obtained for DAMSL and DIT^{++} we may conclude that satisfactory high inter-annotator agreement can be reached when using relatively complex dialogue act annotation schemes. Multidimensional schemes can be applied equally reliably (or even more reliably) as one-dimensional schemes. The usability and reliability of an annotation scheme is not so much a matter of the simplicity of the tag set, but rather of its conceptual clarity, with precise communicative functions definitions and clear annotation guidelines.

4.2 Dialogue units and segmentation

A dialogue act being a unit in the semantic description of communicative behaviour in dialogue, the question arises what stretches of such behaviour are considered as corresponding to dialogue acts. Spoken dialogues are traditionally segmented into *turns*. A turn is defined as[4]:

(10) *stretch of communicative behaviour produced by one speaker, bounded by periods of inactivity of that speaker or by activity of another speaker*
(Allwood, 1992)

Turn boundaries are generally well recognised both by humans and machines. People are able to predict turn endings with high accuracy using semantic, syntactic, pragmatic, prosodic and visual features (Ford & Thompson, 1996; Grosjean & Hirt, 1996; Barkhuysen et al., 2008, among others). It was observed by de Ruiter et al. (2006) that many turn transitions happen without temporal delays because a potential next speaker knows when a turn will end.

Turns, in the sense defined in 10, can be quite lengthy and complicated, and are for most purposes too coarse as the stretches of behaviour to assign communicative functions to. Decomposing a dialogue into turns may suggest that a dialogue can be cut up into sequences of communicative activity of one speaker followed by that of another, but this does not do justice to the complexity of natural communication, especially when more than two participants are involved. In natural communication, where the participants do not only use speech but also facial expressions, gaze direction, head, hand and shoulder gestures, body posture, and nonverbal sounds (laughs, sighs, sucks, chuckles,..), all participants are most of the time performing some communicative activity, as illustrated in Figure 4.1, so the delimitation of turns *by periods of inactivity of a speaker* does not work well. Moreover, it has been found that participants in natural multiparty conversations very often speak overlapping, rather than in sequences of single-speaker turns (see e.g. Campbell, 2008). Taking this into consideration, the notion of a *turn unit* has been introduced and defined as:

(11) *stretch of communicative behaviour produced by one participant which includes the use of speech, and is bounded by periods where that participant is not speaking.*

According to this definition a turn unit is produced by a speaker who may, in addition to speaking, also produce nonverbal communicative behaviour (such as gestures and facial expressions), and turn units produced by different speakers may overlap.

[4] Allwood gives this as a definition of units in spoken dialogue which he calls ‘utterances’.

Speaker	Observed communicative behaviour									
D	words	What's	teletext							
	gaze	averted(table)	personA	personB						
	eyes		narrow							
	posture	working position								
annotation	Feedback	neg. understanding								
	TurnM.	Turn assign to A								
B	words				um	It's	a	British	thing	
	gaze	averted(table)	personD	personA	personD					
	eyes			widen						
	lips			random movements						
	posture	bowing	working position							
annotation	Feedback		pos. attention							
	TurnM.			turn take	turn keep					

Figure 4.1: Example of multimodal communicative behaviour in multiparty dialogue.

Turn units consist of more fine-grained units called *utterances*[5]. Utterances are linguistically defined contiguous stretches of (linguistic) behaviour. Levinson (1983) writes: "An utterance is the issuance of a sentence, a sentence-analogue, or sentence-fragment, in an actual context". For example:[6]

(12) A1: First of all just to kind of make sure that we all know each other
A2: I'm Laura and I'm the project manager

The speaker in A1 introduces the next topic for discussion in a meeting, and in A2 she introduces herself (and the role she will play in the dialogue). A1 and A2 constitute two utterances, together making up a turn unit produced by speaker A.

Segmenting a dialogue into utterances has the advantage of more fine-grained units being annotated, allowing more precise annotation; however, the notion of an utterance as a smaller unit inside a turn does not have a clear definition, and the detection of utterance boundaries is a highly nontrivial task. Syntactic features (e.g. part-of-speech, verb frame boundaries of finite verbs) and prosodic features (e.g. boundary tones, phrase final lengthening, silences, etc.) are often used as indicators of utterance endings (see e.g. Shriberg et al.,1998; Stolcke et al., 2000; Nöth et al., 2002).

The stretches of behaviour that are relevant for interpretation as dialogue acts often coincide with utterances, but they may be discontinuous, may overlap, and may even contain parts of more than one turn. They therefore do not always correspond to utterances, which is why we have introduced the notion of a *functional segment* as a minimal stretch of communicative behaviour that has a communicative function (and possibly more than one).[7] Thus, the units of dialogue that our analysis will be concerned with, are functional segments.

A multidimensional approach to dialogue act annotation naturally leads to abandoning the idea that segmentation should aim at cutting up a dialogue into a linear sequence of stretches of speech, and that one should allow functional units to overlap, to be discontinuous, to include or embed other functional segments, and to spread over multiple turns, leading to a more accurate form of segmentation that other approaches. Moreover, it supports the identification of relevant dialogue segments not only per dimension but also per modality, and the identification of complex multimodal multifunctional segments.

An example of a discontinuous functional segment is (13), where the speaker interrupts his Inform with a Set Question:[8]

(13) Because twenty five Euros for a remote... *how much is that locally in pounds?* is too much money to buy an extra remote or a replacement remote

Overlapping (or embedding) of functional segments is illustrated in (14).[9]

(14) A: What time is *the first train to the airport on Sunday*?
B: *The first train to the airport on Sunday* is at 06:25.

[5] In the literature the term "utterance" is sometimes used to designate everything contributed in a single turn, in the sense of what we call a turn unit, see e.g. Allwood (1992), who uses the term "grammatical unit" for what we call "utterance").

[6] From the AMI meeting corpus - ES2002a.

[7] These stretches are 'minimal' in sense of not being unnecessarily long. The rule here is: do not add material which does not contribute to the communicative function.

[8] From the AMI meeting corpus - ES2002a.

[9] From the OVIS dialogue corpus.

Timeline	Speaker	Dimension						
1195.16 - 1197.56	A1		We're aiming a fairly young market					
		Task	Inform					
1202.88 - 1208.36	B1		Do	you	think	then	we voice should really consider voice recognition	
		Task	Propositional Question					
		Auto-Feedback				Pos.execution A1		
		Turn M.		Turn-Assign to A				
1210.6 - 1211.76	B2		What	What do you think			Craig	
		Task	Set-Question					
		Turn M.	Turn-Take				Turn-Assign to C	
1211.92 - 1217.24	C1		Well	did	you	not say it was the adults that we're going for		
		Auto-Feedback	pos.execution B2 neg.execution B1	Propositional Question to A1				
		Turn M.	Turn-Accept		Turn-Assign to A			

Figure 4.2: Example of multidimensional segmentation.

In this example, B's response as a whole is an answer to A's question, and the repeated question part *The first train to the airport on Sunday* can be viewed as expressing a positive feedback act, displaying B's understanding of A's question.

Example (15)[10] shows that a dialogue act unit may spread over multiple turns. A asks a question, the answer to which consists of a list of items which B communicates one by one.

(15) A: Could you tell me what departure times there are for flights to Frankfurt on Saturday?
B: Certainly. There's a Lufthansa flight in the morning leaving at 08:15,
A: Yes,
B: And a KLM flight at 08:50,
A: Yes,
B: And then a Garoeda flight at 11:45,

The complications of discontinuity, overlap, and spreading over multiple turns can be handled by applying the multidimensional view on communication which is inherent to DIT, and segmenting a given stretch of behaviour 'multidimensionally', in as many ways as there are dimensions in which parts of it have a communicative function. Consider the example in Figure 4.2. Utterance B1 'Do *you* think *then* we should really consider voice recognition' consists of three functional segments in three different dimensions: (1) 'Do you think then we should really consider voice recognition' is a Task Propositional Question; (2) 'you' is a Turn-Assigning act addressed to participant A (also because B directs his gaze to A); and (3) the discourse marker 'then' is a positive Auto-Feedback act at the level of execution related to the previous segment A1.

The multidimensional approach to dialogue act segmentation not only solves various problems concerning the segmentation of dialogue into functional units as illustrated above, but also results in a more accurate analysis expressed in higher scores for automatic dialogue act classification (see Geertzen et al., 2007; Petukhova and Bunt, 2011).

There are still other types of units in dialogue which are relevant for dialogue analysis. For instance, analysing structural relations between several dialogue units we found relations between functional segments and *groups of functional segments*, as the following example shows:[11]

G1: Right. Start off facing North, turn to your left and walk forward, then to your left again. Keep walking forward until you come to the site of a plane crash. Go right roundabout it and turn to your right, so you end up facing North again.
(16) U1: Could you just slow down a bit please?
G2: Sorry.
G3: So you start facing North
U2: Mmhmm

The speaker in U1 is apparently overloaded with the information given in G1, making it hard for him to process these segments successfully. U1 is a negative feedback act relating to the group of four functional segments in G1.

4.3 Relations between dialogue units

In order to analyse what happens in dialogue it is insufficient to only consider its segments in isolation. It is uncontroversial, that discourse modelling requires the consideration of relations

[10] From the Schiphol dialogue corpus.
[11] From the MapTask dialogue corpus.

between semantically or pragmatically relevant units, but the nature, the purpose and the definitions of units in discourse and their relations are the subject of much controversy (see e.g. Hovy, 1990). To the rhetorical relations identified in monologue (e.g explanation, justification, cause,...), dialogue adds relations such as those between a question and an answer, and between an utterance and feedback about its understanding.

Many frameworks for discourse analysis have attempted to capture discourse coherence by integrating all discourse segments into a single structure using discourse relations. Although this has not always been made explicit, the assumption that there is a single "coherence" dimension is strong in many frameworks (Hobbs, 1985a; Mann and Thompson, 1988; Asher and Lascarides, 2003). Grosz and Sidner (1986), followed by Moore and Pollack (1992), on the other hand argued for the interplay between several structures to explain discourse phenomena. Petukhova and Bunt (2009b) have shown that discourse markers are in general multifunctional, requiring a multidimensional approach.

A variety of frameworks for modelling discourse structure have been proposed since Hobbs (1979). While Van Dijk (1979) and Polanyi (1998) have attempted a quasi-syntactic approach, most frameworks are functional in nature and rely on interpretation for deriving a structure of discourse. Relations between discourse segments have in these frameworks been divided into several categories: semantic/ inter-propositional/ ideational/ content-level/ information-level; pragmatic/ intentional/ cognitive/ speech-act; presentational/ structural/ textual; see Hovy et al. (1995) for a discussion of the different categories.

Discourse relations can apply to segments of various size, from syntactic clauses to paragraphs. When considering dialogue, the picture gets even more complicated, with units specific to their interactive nature, such as turn units. Some researchers distinguish between macro-, meso- and micro-levels in discourse structuring (e.g. Nakatani and Traum, 1999; Louwerse and Mitchell, 2003), where the *micro-level* is concerned with relations within a turn unit or within a single utterance; the *meso-level* concerns relations involving complete contributions in Clark's sense (Clark, 1996), typically an initiative and a reactive, corresponding to grounding units; and the *macro-level* is concerned with entire subdialogues, topic structure and elements of a plan-based analysis.

Although often cited as a crucial issue for linguistics and NLP, discourse structure frameworks face the problem of their empirical validation. It is mainly to address this issue that several discourse annotation projects have been undertaken in recent years (Carlson et al., 2001; Wolf and Gibson, 2005; Miltsakaki et al., 2004; Reese et al., 2007; Stede, 2008; Prasad et al., 2008). These ambitious projects share a common goal but differ greatly with regard to their theoretical assumptions. A more generic approach to the analysis of these relations would therefore be of great help for comparing and perhaps combining these accounts.

4.3.1 Functional and feedback dependence relations

Responsive dialogue acts by their very nature depend for their semantic content on the semantic content of the dialogue acts that they respond to. Responsive dialogue acts (also known as 'backwards-looking acts') come in three types:

A acts with a responsive general-purpose communicative function: Answer and its specializations (Confirm, Disconfirm); Agreement, Disagreement and Correction; and Address/Accept/Decline Request, Suggestion, or Offer;

B feedback acts with a dimension-specific communicative function;

C some dialogue acts with dimension-specific functions other than feedback functions, such as Return Greeting, Return Self-Introduction. Accept Apology, Accept Thanking, Return Goodbye; and Turn Accept.

All responsive dialogue acts have a 'functional antecedent', being the dialogue acts that they respond to; those of type A have a semantic content that is co-determined by that of their functional antecedent. This is a relation between two dialogue acts or between a dialogue act and a group of dialogue acts, as in (17).

(17) A1: Can you tell me what time is the first flight in the morning to Munich?
B1: On what day do you want to travel?
A2: Tomorrow.
B2: Tomorrow morning
B3: The first flight that I have is at 7:45.

The dialogue act in B3 is functionally related to the group consisting of the question in A1 and the answer (to B1) in A2, which together are equivalent to a more complete question which B3 answers.

If the meaning of a responsive dialogue act depends on the meaning of a previous dialogue act (or dialogue acts) due to its communicative function, then this dependence is called a *functional dependence relation* (Bunt et al., 2010). More explicitly:

(18) *A functional dependence relation exists between a dialogue act DA_1 and one or more previous dialogue acts $\{DA_2, ..., DA_N\}$ iff the meaning of DA_1 depends on the meaning of $\{DA_2, ..., DA_N\}$ due to the responsive character of DA_1.*

Example (19) shows that the interpretation of A1 clearly depends very much on whether it is an answer to the question B1 or to the question B2, even though A1 would seem a complete, self-contained utterance.

(19) A1: I'm expecting Jan, Alex, Claudia, and David, and maybe Olga and Andrei to come.
B1: Do you know who's coming tonight?
B2: Which of the project members d'you think will be there?

Responsive dialogue acts of type B provide or elicit information about the (perceived) success in processing a segment of communicative behaviour earlier in the dialogue. Such a relation is called a *feedback dependence relation*. This type of relation has been defined in ISO standard 24617-2 as follows:

(20) *A feedback dependence relation is a relation between a feedback act and the stretch of communicative behaviour whose processing the act provides or elicits information about.*

Examples are the relation between U1 and G1 in (16); and between B1 and A1 in (17).

Feedback acts refer explicitly or implicitly to the stretch of dialogue that they provide or elicit information about. This stretch of dialogue forms part of its semantic content. For example, the semantic content of the feedback act in B2 in (17), where the communicative function is Positive Auto-Feedback,[12] has the segments A1 and A2 as its semantic content.

[12] This is the communicative function expressing that the speaker informs the addressee that he (believes that he) understands the utterance that the feedback is about.

In view of this relation between the feedback act and its functional antecedent, one could consider the feedback dependence relation as an instance of he functional dependence relation in the feedback dimension. However, the two relations must be distinguished, since a dialogue act with a functional dependence relation *also*, by implication, has a feedback dependence relation to the segment containing its functional antecedent. For example, an answer implies positive feedback about the speaker's processing of the utterance expressing the question that the answer functionally depends on.

A feedback act does not necessarily refer to a single utterance, but may also relate to a larger stretch of dialogue; even to the entire preceding dialogue, like the global positive feedback expressed by *Okay* just before ending a dialogue. The scope and distance that may be covered by the various kinds of relations in dialogue are analysed in the next section.

A responsive act of type C is related to one or more dialogue acts, like those of type A (functional dependence relation). Such dialogue acts have, however, no or only marginal semantic content, and the meaning of such a dialogue act is concentrated in its communicative function. The semantic interpretation of the dependence relation between such acts is that the dependent dialogue act resolves the reactive pressure created by its antecedent (see Bunt, 1989; 1994).

4.3.2 Rhetorical relations

Rhetorical relations have been proposed as an explanation for the construction of coherence in discourse or at least as crucial modelling tools for capturing this coherence, see e.g. Hobbs (1985a); Mann and Thompson (1988); Sanders et al. (1992); Asher and Lascarides (2003). The idea is that two text segments or sentences in written discourse, or two segments or utterances in dialogue, are linked together by means of certain relations, for which various terms have been used such as 'rhetorical relations', 'coherence relations', or 'discourse relations'.

Their study can be traced back to the Antiquity, with a continuous attention from rhetorics over the centuries, but the way they have been used recently in AI and NLP probably comes from Hobbs' seminal work in this area (Hobbs, 1979). Since then a range of taxonomies have been proposed in the literature to define relations in discourse. The well-known set of relations and their organization proposed by Mann and Thompson (1988), forming the core of Rhetorical Structure Theory, consists of 23 relations. Hovy and Maier (1995) studied approximately 30 alternative proposals and proposed a hierarchical taxonomy of approximately 70 relations.

Some rhetorical relations, such as *Explanation*, *Justification*, and *Cause* are clearly semantic, whereas others, like *First*, *Second*,..., *Finally*; and *Concluding* are more presentational in nature. The occurrence of truly semantic rhetorical relations is illustrated in example (21) from the AMI corpus, where participant A talks about remote controls:

(21) A1: You keep losing them
A2: It slips behind the couch or it's kicked under the table

The events described in these sentences are semantically related by *Cause* relations: *Cause* (slipped; keep_loosing) and *Cause* (kicked; keep_loosing). In cases like this the two sentences are related through a rhetorical relation between the events they contain. We use the term *'inter-propositional relation'* for rhetorical relations between the semantic contents of two dialogue acts, irrespective of whether these semantic contents are in fact propositions; in particular, they may very well be events (or more generally 'eventualities', see Bach, 1981).

Contrary to what is sometimes believed, semantic rhetorical relations are not always relations between events (or 'eventualities'). Consider the following example, where A and B discuss the use of remote controls:

(22) A: You keep losing them
B: That's because they don't have a fixed location

The 'event' in the second utterance (*having a fixed location*) does not cause the *losing* event in the first utterance; on the contrary, the second utterance says that the fact that *no* having-a-fixed-location event occurs is the cause of the *losing*. Saying that a certain type of event does *not* occur is not describing any event, but expressing a *proposition* (about that type of event). This means that the causal connection between the two utterances is not between two events, but between the *proposition* made in the second utterance and the event in the first utterance.[13]

Rhetorical relations between dialogue utterances do not necessarily relate the *semantic contents* of the dialogue acts that they contain, but may also relate the *dialogue acts* as such, taking both their semantic contents and their communicative functions into account. The following examples[14] illustrate this:

(23) A1: Where would you physically position the buttons?
A2: I think that has some impact on many things

(24) B1: I'm afraid we have no time for that.
B2: We're supposed to finish this before twelve.

Utterance A2 in (23) encodes an Inform act which has a *Motivation* relation to the Question act encoded in A1; it tells the addressees what motivated A to ask the question A1 with this particular semantic content. In (24) utterance B2 encodes an Inform act which has an *Explanation* relation to the Decline Request act in B1.

4.3.3 Scope and distance

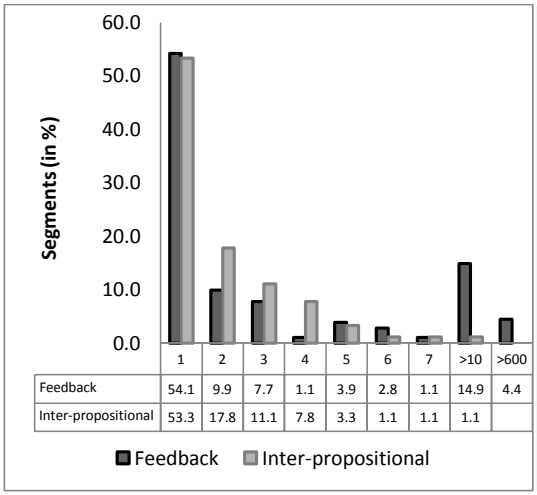
While a feedback dependence relation can target a functional segment, a dialogue act, a turn unit, or a group of those, functional dependence and rhetorical relations are grounded in meaning and follow more restricted patterns of linking. We investigated the linking patterns of the different types of relations for two corpora of annotated dialogues, the AMI meeting corpus and a French two-party route explanation dialogues collected at the University of Toulouse.[15]

For analysing these patterns it is helpful to look at the *scope* and *distance* covered by a relation. We define scope as follows:

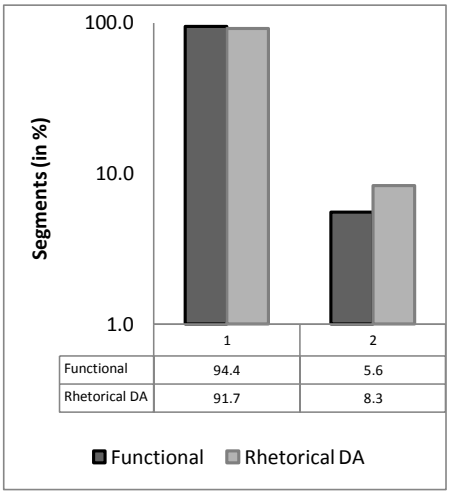
(25) *the scope of a discourse relation is the number of functional segments (the 'target') that a given segment (the 'source') is related to.*

Calculation of the distance between related functional segments in dialogue is not a trivial task and deserves some discussion. The distance between two segments can be calculated *textually*, e.g. as the number of intervening constituents between a pair of constituents under consideration, or as the number of intervening tokens; and *topologically*, as the length of the

[13] It could in fact be argued that the first utterance also contains a proposition, rather than describing an event.
[14] From the AMI meeting corpus - ES2002a.
[15] For more information see Muller and Prévot (2003) and `http://crdo.fr/crdo000742`



	1	2	3	4	5	6	7	>10	>600
Feedback	54.1	9.9	7.7	1.1	3.9	2.8	1.1	14.9	4.4
Inter-propositional	53.3	17.8	11.1	7.8	3.3	1.1	1.1	1.1	



	1	2
Functional	94.4	5.6
Rhetorical DA	91.7	8.3

Figure 4.3: Scope of feedback dependence, functional dependence and rhetorical relations in the AMI data.

shortest path in a graph representation (e.g. a SDRS, see Afantenos and Asher, 2010). Since in this study we did not construct any tree or graph representations for the various kinds of relations we distinguished, we considered the textual calculation of distance between related segments. In dialogue, the most plausible unit for measuring distance is the functional segment, but simple count of intervening functional segments is not possible, because of the following complications:

- Spontaneous speech includes self-corrections, retractions and restarts that have a communicative function of their own and are considered as functional segments. Speech errors and flaws like reparanda (segment parts that are retracted or corrected by the same speaker) do not have any communicative function on their own;
- Functional segments may be discontinuous and may be interrupted by more substantial segments than repairs and restarts, e.g. 'Because twenty five Euros for a remote... *how much is that locally in pounds?* is too much money to buy an extra remote or a replacement remote';[16]
- Functional segments may be spread over more than one turn, e.g. A: Well we can chat away for ... um... for five minutes or so I think at... *B: Mm-hmm* ... at most;[17]
- Functional segments may overlap, e.g. U: What time is the first train to the airport on Sunday? S: *The first train to the airport on Sunday* is at 6.17, where the part in italics forms part of an answer to U's question, but also has a feedback function, displaying what S has heard;
- In multi-party interaction multiple participants may react to the speaker's previous contribution and may do this simultaneously, with some overlap or consecutively, e.g.

(26) A: Do you have anything to add?
B: *No*
C: *No*

[16] From the AMI meeting corpus - ES2002a.
[17] From the AMI meeting corpus - ES2002a.

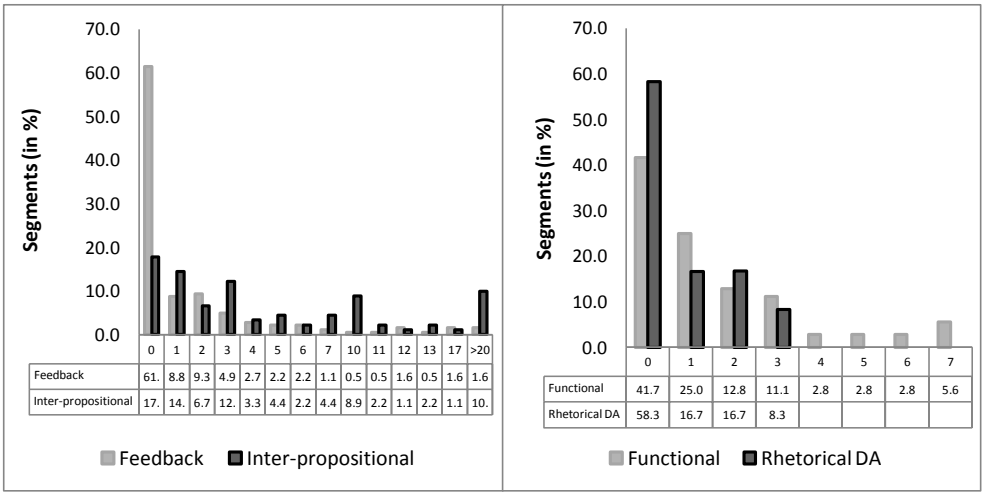


Figure 4.4: Distance of feedback dependence, functional dependence and rhetorical relations in AMI data.

These dialogue-specific phenomena should be taken into account while calculating the distance between related functional segments. All segments were ordered by their start time. Given two non-overlapping segments A and B, with begin(B) $\geq$ end(A) (i.e. B starts when or after A has ended) a segment C is counted as intervening between A and B if and only if C starts later than A and before B. (In that case, C must contain some material that occurs after A has started and before B has started) We thus define:

(27) *A segment C intervenes between the segments A and B iff begin(C) $>$ begin (A) and begin(C) $<$ begin(B).*

Moreover, if C and D are two intervening segments with the same begin- and end points, with the same communicative function(s) and with identical semantic contents (but contributed by different speakers), (cf. (26)), then they are counted only once, and if an intervening functional segment E is a sub-segment of a larger intervening segment K produced by the same speaker, then only the larger segment is counted.

If A and B are overlapping or consecutive segments, i.e. begin(B) $\leq$ end(A), we stipulate their distance to be zero. Hence we use the following definition of distance:

(28) *The distance between two non-overlapping segments A and B, with begin(B) $>$ end(A) equals the number of intervening functional segments* minus *the number of co-occurring intervening functional segments with identical wording and interpretation (produced by different speakers)* and minus *the number of sub-segments of intervening functional segments produced by the same speaker.*

Moreover, in order to deal with the complications mentioned above, we removed all reparanda from the data, e.g. 'This is the kick-off meeting for our our project' became 'This is the kick-off meeting for our project'; and we merged functional segments that were spread over multiple turn units.

Figure 4.3 shows the scope and Figure 4.4 the distance involved in functional and feedback dependence relations, and for inter-propositional relations and rhetorical relations between dialogue acts, as found in the AMI corpus. Our analyses show that different relations exhibit different patterns. A functional dependence relation normally has a narrow scope (1-2 functional segments), and units related by this type of relation tend to be close to each other in discourse (distance 0-2), except in the case of discourse pop-up units (see below). Feedback dependence relations as a rule have either very narrow (1-2) or very wide scope (5-10); long-distance attachments are rare. Feedback acts can target all types of dialogue units that we have defined: other dialogue acts, turn units, functional segments, as well as groups of those. Rhetorical inter-propositional relations often have narrow scope but the related segments may be some distance away from each other. Rhetorical relations between dialogue acts are generally characterized by narrow scope and short distance, but some rhetorical relations (like *Recap, Conclude*) link a dialogue act or a dialogue act group to one or more dialogue act groups, having a wide scope.

Four types of attachment in terms of distance and scope can be distinguished for the way in which discourse relations connect a source segment to other units in dialogue:

1. Last segment: A relation links the source segment to the previous functional segment: scope is 1, distance is 0.
2. Local attachment: The source segment is related to several previous segments. The scope of each relation is 1; at least one of the relations has distance 0, and at least one has distance greater than 0. For example, the next step of a narration introduces a contrast with the preceding segment, while elaborating an earlier segment.
3. Local wide scope attachment: The relation targets a group of segments. The scope is greater than 1, the distance is 0. This is common with relations such as *Recap, Summarize, Conclude*.
4. Discourse pop-up: The source segment is related to an earlier functional segment or to a group of functional segments: the distance is greater than 1, the scope is 1 or greater.

Attachments of type 1 occur frequently (29.8% of all attachments in the AMI corpus). For example:[18]

(29) D1: Cost like production cost is twelve fifty or retail like on the shelf?
B1: Our sale anyway
B2: Because its probably up to the retailer to sell it for whatever price they want

Segment B1 is an Answer to the Choice Question in D1, and segment B2 provides a Justification for the Answer in B1.

Attachments of type 2 are more complicated. Such attachments are frequently observed in the AMI data, accounting for 41.5% of all attachments. For example:[19]

(30) D1: Now remote controls are better
D2: But actually still kind of like a massive junky thing [Contrast:D1]
B1: Still feels primitive [Elaboration:D2;Contrast:D1]

[18] From the AMI meeting corpus - ES2002a.
[19] From the AMI meeting corpus - ES2002a.

The fact that the related segments are produced by different speakers has the consequence that they exhibit not only rhetorical but also feedback dependence relations by implication, e.g. the expression of Agreement in B1 entails positive feedback on understanding D2.

Local wide scope attachment is frequently observed for feedback dependence relations. Very often feedback is provided not to a single functional segment but to the discourse unit that is concerned with one of the dialogue sub-tasks or topics. This occurs frequently in multi-party dialogues (19.2% in the AMI meetings). Both positive and negative feedback are observed to sometimes have local wide scope attachment. For example:[20]

(31) B1: We're gonna be selling this remote control for twenty five euro
B2: And we're aiming to make fifty million euro [Narration:B1]
B3: So we're gonna be selling this on an international scale [Elaborate:B1&B2]
B4: And we don't want it to cost more than twelve fifty euros [Narration:B3]
D1: Okay [PositiveFB:B1-B4]
B5: So fifty percent of the selling price [Conclude:B1&B4]
D2: Can we go over that again [NegativeFB:B1-B5]

Muller and Prévot (2003) have shown that in French route explanation dialogues, *voilà (that's it)* is a marker of closure, thus being some kind of wide-scope feedback (type 3) preparing a discourse pop-up of type 4. For example:[21]

(32) G13: Hop hop hop Esquirol tu continues tout droit(hop hop hop Esquirol continue straight)
G14: Y'a le Classico (there is the Classico)
R15: Euh (uh)
G16: T'as pas l'air branché trop bars (you do not seem to be into bars)
R17: Euh non (uh no)
R18: Mais je connais pas très bien Toulouse (but I don't know Toulouse very well)
G19: Ah ouais d'accord (ah yeah ok)
G20: Donc Les Carmes tu vois ou c'est? (so Les Carmes, you know where it is?)
F21: Oui (yes)
G22: Bon ben voilà. (well that's it)
G23: Donc là tu continues sur sur cette rue (so there you continue on this street)
G24: Et tu arrives aux Carmes (and you get to Les Carmes)

Segment G22 concludes and closes discourse unit [G14-G21], and a Continuation/Narration relates G13 to G23 (discourse pop-up attachment).

Many discourse markers, which have been studied for their semantic contribution and for their role in dialogue structuring, are good indicators of various kinds of discourse attachment. Most connectives (*then, but, therefore*) connect functional segments with attachment of type 1 or 2. Enumerative markers such as (*First, Then, Finally*) can introduce macro-structures resulting in both long-distance and local wide-scope attachment, since usually the entire discourse unit that contains these markers is rhetorically related to another discourse unit.

Figure 4.9 in Section 4.6 summarizes our qualitative findings in the form of an ISO-style metamodel (cf. Bunt et al., 2010) containing the various kinds of units in dialogue and the possible relations between them.

In future, more comprehensive studies of the properties of discourse relations in dialogue, such as their scope and distance, it may be useful to distinguish between the semantic and presentational dimensions of rhetorical relations.

[20] From the AMI meeting corpus - ES2002a.
[21] From the French route navigation corpus.

4.4 Communicative function qualification

Participants in a dialogue do not just exchange information by simple statements, direct questions and clear-cut answers. They may be less straightforward in expressing their communicative intentions, formulating a question indirectly or accepting a request conditionally. They often indicate their attitude toward their communicative partners, toward what they are saying, or toward things that they intend to do. They emphasize, express doubts, criticize, show interest, and so on. All this can be signalled in various ways, e.g. by using verbal indicators like modals, by intonation and by utilizing body language and facial expressions. Approaches to the analysis, annotation, or computational modelling of dialogue behaviour struggle with these phenomena. This is especially true for attempts to annotate spoken and multimodal dialogue with information about the communicative actions ('dialogue acts') that the participants perform.

The analysis of existing well-known dialogue act annotation schemes showed that virtually every dialogue act taxonomy fails to capture the nuances in the performance of communicative actions relating to certainty, conditionality and sentiment. For example:

(33) 1. A: Would you like to have some coffee?
2. B: Only if you have it ready.
3. B: Coffee could be nice, but what time is it now?

Response 2 in (33) can be characterized as *conditional acceptance of offer* and response 3 as *uncertain acceptance of offer*.

Some dialogue act taxonomies pay attention to these phenomena. For instance, DAMSL and DAMSL-based schemes like SWBD-DAMSL, MRDA and Coconut distinguish such functions as Reject-Part or Accept-Part. To address uncertainty DIT^{++} Release 4 (see `http://dit.uvt.nl/dit4`) includes the communicative functions Uncertain Answer, Uncertain Agreement, Uncertain Disagreement, Uncertain Confirm and Uncertain Disconfirm, and conditionality is captured by the functions Indirect Request, Indirect Set Question, etc. This is not really a way to go, however, since for example an Accept Offer can be uncertain and indirect at the same time, and expressed with a certain sentiment as well, so this would lead to an undesirable growth of the tag set, undermining its transparency. Instead, we propose to add a set of *qualifiers* that can be attached to a communicative function in order to describe the speaker's behaviour more accurately.

A qualifier is an additional element in the description of dialogue acts. Semantically, qualifiers describe and provide more accurate definitional meaning for another element. Communicative function qualifiers do not change but specify more precisely the way the act's semantic content changes the addressee's information state, e.g. by expressing the strength or weakness of certain assumptions and beliefs, or the physical and emotional abilities and state of a dialogue participant. In other words, qualifiers provide a more detailed description of the speaker's intention.

Most existing dialogue act taxonomies consider only two possible responses to an offer, a suggestion, or a request: accepting it and rejecting it (additionally, DAMSL defines Hold). However, people often respond in a less straightforward way, e.g. accepting the offer conditionally or with a certain modality. Consider the following example:

(34) A: Would you like to have some coffee?
1. B: I'm not sure I want any.
2. B: Maybe later?

3. B: Yes, I definitely need some.
4. B: Yes, please, if you don't mind to bring it for me.
5. B: Coffee? At midnight?
6. B: Yes, a strong one, or I fall asleep

Response 1 can be seen as *uncertain* acceptance/rejection of A's offer; in response 2 acceptance is expressed by communicating probability; response 3 can be characterised as *c*ertain acceptance; response 4 is a *conditional* acceptance; and in response 5 the speaker signals surprise, without clearly accepting or rejecting the offer. In response 6 the speaker believes that it is appropriate to provide additional or more detailed information to the acceptance of the A's offer, and he also explains why he accepted it.

To summarize, at least three categories of qualifiers, *certainty*, *conditionality* and *sentiment*, deserve to be analysed in more detail. In (Petukhova and Bunt (2010b)) a category of qualifiers for *partiality* was proposed as well. After further analysis, however, it was concluded that partiality is not a communicative function qualifier, but rather a property of semantic content. Answers are commonly non-exhaustive, i.e. partial, as in the following examples:

(35) A: Do you know who'll be coming tonight?
1. B: Peter, Alice, and Bert will come for sure.
2. B: I heard that Tom and Anne will not come.
3. B: I have a hunch that Mary will not come.

The responses 1, 2 and 3 in (35) all constitute partial answers. Response 3 is also uncertain.

Thus, (non)-exhaustiveness is rather a property of the relation between a dialogue act and ints functional antecedent.

4.4.1 Qualifier definitions and uses

Certainty

Modality is generally seen as a category of linguistic meaning which is concerned with expressions of certainty. Mindt (1998) distinguishes 17 modalities: (i) possibility/high probability, (ii) certainty/prediction, (iii) ability, (iv) hypothetical event/result, (v) habit, (vi) inference/deduction, (vii) obligation, (viii) advisability/desirability, (ix) volition/intention, (x) intention, (xi) politeness/downtoning, (xii) consent, (xiii) state in the past, (xiv) permission, (xv) courage, (xvi) regulation/prescription, and (xvii) disrespect/insolence. Leech (1971) proposed to differentiate between 11 modal meanings: (i) possibility (theoretical, factual), (ii) ability, (iii) permission, (iv) exclamatory wish, (v) obligation/requirement, (vi) rules and regulations, (vii) local necessity, (viii) prediction/predictability, (ix) willingness (weak volition), (x) intention (intermediate volition), and (xi) insistence (strong volition). The most widely used division of the modal domain distinguishes between (i) alethic, (ii) deontic, (iii) dynamic and (iv) epistemic.

Alethic modality is concerned with degrees of certainty of the truth of a proposition; this is a category of modal logic for which it is not easy to find examples in natural language. *Deontic* modality is concerned with what is possible, necessary, permissible or obligatory according to law or social and moral obligations, and refers to actions and events. *Dynamic* modality refers to physical necessity or possibility and is concerned with expressions of ability, power, futurity, prediction and habit. This modality is applicable to propositions as well as to actions.

Epistemic modality is concerned with what is possible given what is known and what evidence is available. Epistemic modals form an interesting category to be studied "*because their*

semantics is bound up both with our information about the world and with how that information changes as we share what we know" (von Fintel and Gillies, 2007). The semantics of epistemically modalized utterances, which is context-dependent, is still under debate. Von Fintel and Gillies suggest that utterances with epistemic modals are used to perform more than one speech act. For example:

(36) There might have been a mistake in calculation

They argue that (36) is an *assertion* and an *explanation*. This analysis is in our opinion not correct, because by the assertion the speaker wants to make something known to the addressee, and explanation always subsumes assertion; in other words, making an assertion plus an explanation is semantically the same as an explanation *per se*.

Potts (2003) and Swanson (forthcoming) propose to treat epistemic modals as 'speech act modifiers'. Swanson suggests that an unmodalized sentence has to be interpreted as an assertion and a modalized sentence as 'assertion with tempered force' which could have the appropriate kind of context change potential. This approach is potentially promising. Epistemically modalized utterances may be considered as having a ***qualified*** communicative function.

Epistemic modal qualifiers are concerned with expressions of the speaker's degree of certainty regarding the validity of a proposition. For example:[22]

(37) 1. I think that for the next meeting we have market data
2. I guess generic remote is what we're aiming for

In the utterances in (37) the speaker *weakly* believes that the propositions are true.

Uncertainty is often communicated through expressions of 'probability'. For example:[23]

(38) It will probably be sold separately

Our corpus analysis shows that dialogue participants often express assessments of the validity of their propositions. About 47% of all utterances are modalized (34.5% uncertain, 12.6% certain). A degree of certainty can be expressed verbally as well as nonverbally. Table 4.3 gives an overview of observed expressions of (un-)certainty.

Conditionality

Conditional qualifiers refer to the possibility (with respect to ability and power), necessity or volition of performing actions, and can therefore be attached to action-discussion functions. Consider the following examples:[24]

(39) 1. If you're ready, maybe you make your presentation
2. I can do this for you if you like
3. I'll send you an e-mail if you give me your address
4. If we want a few more buttons maybe we could have them in a little charging station

[22] From the AMI pilot meetings.
[23] From the AMI meeting corpus - ES2002a.
[24] From the AMI meeting corpus.

Table 4.3: Expressions of modality.

Modality	Verbal expressions	Vocal /prosody	Gaze direction	Head movement	Facial expression	Gesture	Posture orientation
Uncertainty	may (not) might (not) could (not) should (not) probably(not) (un)likely maybe(not) 'not sure' 'you know?' 'I guess', etc.	high sd in pitch; voice breaks; jitter; shimmer; filled/ unfilled pauses;	aversion redirection involuntary eye movements	waggles	lip compressed; lip-pout; biting/liking; lowering eyebrows; constricting forehead muscles	adaptors, e.g. self-touch; shoulder shrug	posture shift
Certainty	shall will(not) can(not) would(not) must(not) certainly(not) definitely(not)	low standard deviation in pitch; no pauses no restarts	direct eye contact;	head nod (for emphasis)	thin lips; pushing up the chin boss; widely open eyes;	beat gestures	leaning forward /to addressee

Utterance 1 in (39) is a *conditional request*; utterance 2 a *conditional offer*; utterance 3 a *conditional promise*, and utterance 4 a *conditional suggestion*.

Some communicative functions are inherently conditional. For instance, a *request* to do X can be seen as a *conditional instruct* to do X (the condition being that the addressee agrees to do X), and an *offer* to do X can be viewed as a *conditional promise* to do X (the condition that the addressee accepts the offer). Indirect requests are conditional on the addressee's consent or ability to perform the requested action. For example, in (40) the speaker asks the addressee to explain something on the condition that he is able to do so:[25]

(40) Can you explain this?

Responses to action-discussion functions can also be conditionally qualified:[26]

(41) A: Maybe we could have something like a touch screen
1. B: I don't think so, unless it doesn't take lots of space
2. B: If we can do that, great

(42) A: Can we just go over that again
1. B: Just very quickly. I have to hurry you on here
2. B: We have no time, unless you make it very quickly

(43) A: I can make buttons larger
1. B: If it's possible, why not
2. B: No, only if we want basic things to be visible

Response 1 in (41) can be seen as *conditional declining of a suggestion*; response 2, by contrast, is a *conditional acceptance of a suggestion*. Similarly, response 1 in (42) expresses a *conditional acceptance of a request* and response 2 a *conditional declining*. Response 1 in (43) is a *conditional acceptance of an offer* and response 2 a *conditional declining*.

Our corpus analysis shows that about 2.6% of all utterances are conditional. The conditionality is mostly articulated by conditional clauses with 'if' and 'unless', or phrases consisting of 'if' followed by an adjective, e.g. 'if necessary', 'if possible'.

Sentiment

Sentiment is a broad category of qualifiers concerned with the speaker's attitude and emotional state.

A dialogue participant may express his attitude towards the addressee(-s), or towards the content of what he is saying. Attitudes can be divided into positive and negative. Positive attitudes towards the addressee can be articulated e.g. by being polite or friendly. Positive attitudinal expressions include compliments and expressions of appreciation of the addressee's actions, sympathy with the addressee, as well as downplaying his mistakes. Negative attitudes can be expressed by the speaker being offensive or impolite.

Speaker attitudes can also be derived from modality and conditionality. For instance, by formulating a claim with some degree of uncertainty the speaker often wants to appear less assertive, or to 'save the addressee's face'. Conditional acts are often perceived as more polite than unconditional ones, as in indirectly formulated requests.

[25] From the AMI meeting corpus.
[26] Modified utterances from the AMI meeting corpus.

Table 4.4: Facial expressions corresponding with Ekman's six basic emotions.

Emotion	Facial expression						
	Forehead	Eyebrows	Eyes	Nose	Cheeks	Lips/Mouth	Chin
Anger	wrinkled	lowered; pulled together	lower eyelids tensed and straightened			lips tensed; lips pressed together	pushing up of the chin boss
Disgust		pulled down	lower eyelid tensed upper eyelids raised; opening narrowed	wrinkled		upper lip drawn up; lips pressed together; mouth open	
Fear		raised straight up	eyelids raised up			lip corners pulled; lips stretched; mouth open	jaw dropped
Happy			eyelids narrowed; eye corners wrinkled		outer, upper area of the cheeks raised	lip corners raised	
Sad	wrinkled	pulled together and raised in the center of forehead	narrowed		raised cheeks	lip corners pulled down; lips stretched; lip corners downturned	chin boss pushed up
Surprise	wrinkled	raised straight up	upper eyelids raised (slightly to extremely)			mouth opened; lips tensed or relaxed	jaw drop

Attitudes towards the content of an utterance can be expressed by emphasizing its importance, and by positive or negative evaluation of partner's previous related contributions. To stress the importance the speaker can use expressions like 'above all', 'actually', 'believe me', 'by all means', 'indeed', 'really', 'surely', etc. Speakers may use their bodies to indicate that what they are saying deserves special attention, e.g. hand beat gestures are known to accompany new important information, and eyebrow movements may indicate where the focus of the addressee's attention should be positioned.

The evaluation of partner's utterances may be both attitudinally and emotionally loaded. The attitudinal aspect is more related to mental or cognitive processing, while the emotional aspect refers to the feelings the message evokes.

Emotions can be also evoked by the addressee's behaviour. One of the best known taxonomies of emotions is in Ekman's is pioneering work (Ekman, 1972), which distinguishes 6 basic emotions: anger, disgust, fear, happiness, sadness and surprise. In his later work, Ekman (1999) expanded his taxonomy and added 11 more emotions: amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure and shame. Some emotions can be modified to form complex emotions.

In recent years several schemes for annotating emotion-related states have become available. Craggs and McGee Wood (2004) distinguish along with basic emotions like happiness, sadness also affection, dislike, and misery. Laskowski and Burger (2005) distinguish between the description of behaviour and feelings, noting that annotators tend to describe how people behave rather than how they feel. To label emotions in a participant's behaviour they have labels like objecting, protesting, etc. Feelings are analysed in terms of valence: positive, negative and neutral.

In support of the design of the AMI annotation scheme (AMI Consortium, 2005a) experiments were carried out by Ordelman and Heylen (2005) where subjects were provided with a list of 243 terms describing emotions and were asked to select the 20 most frequent ones occurring in AMI meetings. In this way 26 emotional and attitudinal terms were selected. After annotation experiments using these terms, the following emotional and attitudinal states were defined in the AMI scheme: neutral, curious, amused, distracted, bored, confused, uncertain, surprised, frustrated, decisive, disbelief, dominant, defensive and supportive. Inter-annotator agreement in terms of Krippendorff alpha was found to vary from 0.061 to 0.443 (Reidsma et al., 2006).

To summarize, several taxonomies label emotional and attitudinal phenomena in dialogue with different levels of granularity: coarse (positive, negative and neutral); medium (basic emotions comparable to Ekman's 6 emotions), and fine (labels for specific emotions like misery, annoyed, worry, etc., specific attitudes like criticism, impatient, agreeable, serious, curious, etc.). This suggests that it is sensible to leave this category open, leaving the choice of specific sentiment qualifiers to different applications and tasks.

4.4.2 Qualifier recognition

We assessed the recognizability of the qualifiers defined above in a series of annotation experiments, and measured the inter-annotator agreement in terms of standard kappa. The annotation task was to assign the qualifier values discussed above to the functional segments from the selected pre-annotated dialogue fragments from the AMI (105 segments) and TRAINS (53 segments) corpora.

The experiments were performed by naive annotators, four undergraduate students without

linguistic training that have had an introductory session to get familiar with the dialogue data and the tag set. They had been introduced to the annotation scheme and the underlying theory while participating in a course on pragmatics. During the course they had been exposed to approximately three hours of lecturing and a few small annotation exercises. All annotators accomplished the task individually, having received the materials (transcriptions, sound and video) in electronic form. Time was not limited; annotators were allowed to spend as much time as they needed to perform the task, e.g. listen to the audio recordings as many times as they like. For the AMI fragments annotators considered all three qualifier types; for the TRAINS data sentiment qualifiers were left out, since dialogues are not video recorded.

Table 4.5 lists the qualifier attributes and values used in the experiments, indicating in the rightmost column the categories of communicative functions to which they may be attached. These names are identical to the ones that are part of the ISO annotation standard ISO 24627-2, except that for sentiment qualifiers three values were used: 'neutral', 'positive' and 'negative'.

Table 4.5: Function qualifier attributes, values, and function categories.

qualifier attribute	qualifier values	communicative function category
certainty	neutral,uncertain,certain	information-providing functions
conditionality	conditional, unconditional	action-discussion functions
sentiment	neutral, positive, negative	information-providing functions; feedback functions

Table 4.6 shows that there are no systematic differences between annotators in assigning values for qualifier tags. They achieved moderate agreement on labelling certainty for the AMI data; the agreement on this category when labelling TRAINS dialogues is substantial. This can be explained by the fact that AMI dialogues are more difficult to annotate for naive annotators. AMI meetings are much more complex and show a great variety of phenomena and mechanisms that express modality.

The best recognized category is conditionality. Annotators were able to achieve substantial $0.6 < \kappa < 0.8$ to near perfect $\kappa > 0.8$ agreement.

Table 4.6: Cohen's kappa scores for two sets of rating experiments per annotators pair.

Annotators pair	AMI data			TRAINS data	
	Certainty	Conditionality	Sentiment	Certainty	Conditionality
1 & 2	0.49	0.79	0.70	0.64	0.88
1 & 3	0.48	0.64	0.66	0.70	0.73
1 & 4	0.42	0.65	0.25	0.64	0.93
2 & 3	0.47	0.85	0.60	0.68	0.64
2 & 4	0.35	0.79	0.36	0.71	0.88
3 & 4	0.38	0.65	0.30	0.75	0.73

Annotators experienced more difficulties assigning sentiment ($0.25 < \kappa < 0.7$) and uncertainty ($0.35 < \kappa < 0.75$) qualifiers than when labelling conditionality. Some annotators assign sentiment and uncertainty qualifiers to all communicative functions, using 'neutral' and 'positive' and 'negative' values for sentiment, and 'neutral', 'certain' and 'uncertain' values for uncertainty. Others assigned qualifiers only to some communicative functions where they thought either positive or negative sentiment, or certainty or uncertainty was expressed, and considered

the 'neutral' value as default. Agreement for certainty qualifiers is higher for the TRAINS dialogues, maybe because for identifying qualified segments annotators needed to deal only with the speech modality, and additional visual information for AMI dialogues led to disagreement in interpretation. Thus, recognition of (un-)certainty causes some disagreements for complex data (consistent moderate inter-annotator agreement). In general, emotions and attitudes are difficult to recognize if no or limited information about the dialogue participants is available, and perception of emotions often results in quite subjective judgments, see Reidsma (2008). In the AMI meetings, it should also be noted that participants are reserved in showing emotions, and nearly only show positive attitudes. The problematic recognition of sentiment qualifiers (ranging from poor to substantial agreement) may be interpreted as support for the view that sentiment values are best specified for a given task, where they are most relevant, rather than in a context-independent fashion.

4.5 Coding dialogue data with dialogue acts

4.5.1 Dialogue corpus material

In our empirical studies we used data from six different corpora: three corpora in English and three corpora for Dutch. Table 4.7 gives an overview of corpus data with information of type, language, size and purpose of use in what type of study.

The **AMI corpus**[27] contains human-human multi-party interactions in English. Meeting participants (normally 4) play different roles in a fictitious design team that takes a new project from kick-off to completion over the course of a day. The *AMI corpus* contains manually produced orthographic transcriptions for each individual speaker, including word-level timings that have been derived using a speech recogniser in forced alignment mode. The meetings are video-recorded and each dialogue is also provided with sound files (for our analysis we used recordings made with close-talking microphones to eliminate noise).

The **IFA Dialog Video corpus**[28] contains two-party interactions in Dutch. IFA dialogues are informal spontaneous conversation of two participants about food, holidays, etc.

The **OVIS**[29] dialogues are task-oriented human-computer dialogues where the user is expected to obtain information about train connections and schedules. We were provided with transcribed speech and some prosodic information for user utterances.

The **DIAMOND** corpus (Geertzen et al., 2004) contains human-machine and human-human Dutch dialogues that have an assistance-seeking nature. The dialogues were video-recorded in a setting where the subject could communicate with a help desk employee using an acoustic channel and ask for explanations on how to configure and operate a fax machine. The dialogues were orthographically transcribed, and the human-human subset of the corpus was selected for our studies.

The **Schiphol** (Amsterdam Airport) Information Office dialogues (Prüst et al., 1984 and Cramer, 1985) are information-seeking dialogues where an assistant is requested to provide a client information concerning airport activities and facilities (e.g. timetable, security, etc.).

[27] Augmented Multi-party Interaction, for more information visit `http://www.amiproject.org/`

[28] For more information and downloads visit `http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFADVcorpus/`

[29] Openbare Vervoer Informatie Systeem (Public Transport Information System), see `http://www.let.rug.nl/\~vannoord/Ovis/`

Table 4.7: Corpora used in studies.

Corpus	Type	Language	No. dialogues	Duration	No. turns	No. segments	Used in
AMI	multi-party human-human video-/ sound recordings	English American/British native speakers*	5	2h45min	1,339	6,238	study of (non-)verbal behaviour classification experiments
IFADV	two-party human-human video-/ sound recordings	Dutch native speakers	5	1h15min	na	na	study of nonverbal behaviour
MapTask	two-party human-human sound recordings	English native speakers	6	26 min	1442	2589	study of verbal behaviour/classification
OVIS	two-party human-computer	Dutch native speakers	108	3h10min	3,738	5,242	classification experiments
DIAMOND	two-party human-human	Dutch native speakers	4	21min	299	1,408	study of verbal behaviour
TRAINS	two-party human-human	English native speakers	5	11min	117	349	study of verbal behaviour
Schiphol	two-party human-human	Dutch native speakers	6	na		202	study of verbal behaviour

*The AMI corpus contains mostly non-native speakers; however, for our studies we selected dialogues with participants who are English native speakers, although they speak different dialects, e.g. American and British English, and with different accents, e.g. Irish and Scottish.

The **MapTask**[30] dialogues are so-called instructing dialogues where one participant plays the role of an instruction-giver while another participant, the instruction-follower, navigates through the map. The MapTask corpus contains orthographic transcriptions for each individual speaker, including word-level timings.

The **TRAINS**[31] dialogues are information-seeking dialogues where an information office assistant helps a client in planning optimal train transportation of cargo. The TRAINS corpus contains orthographic transcriptions for each speaker.

4.5.2 DIT^{++} multidimensional dialogue act taxonomy

In the DIT^{++} taxonomy communicative functions are organised in a 10-dimensional hierarchical taxonomy. Figure 4.5 shows the DIT 10-dimensional hierarchy, where dimensions are in gray filled boxes. A top-level distinction is made between communicative actions advancing the *underlying task*, such as instructions, questions, and answers, and actions that manage the *communicative task*, such as acknowledgments, attention signals, self-corrections, and turn-taking signals. These actions are called *task-related* dialogue acts and *dialogue control* acts, respectively (see Bunt, 1994).

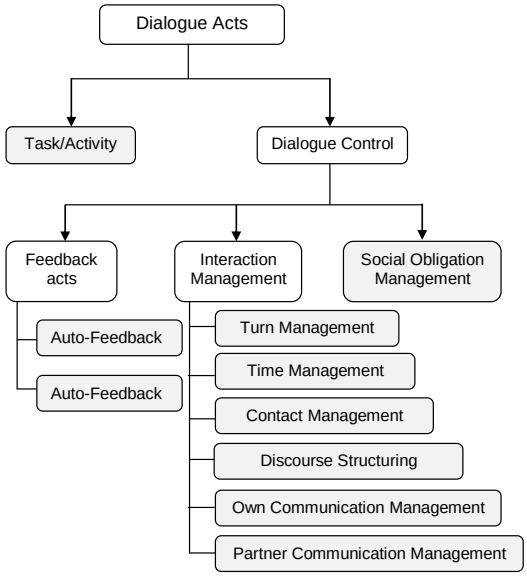


Figure 4.5: DIT^{++} 10-dimensional hierarchy.

The *Task* dimension is formed by dialogue acts intended to advance the underlying task. Task-related acts are only a relatively small part of what happens in natural conversation, however (see e.g. Chapter 3, Section 3.5, Table 3.1 for distribution of dialogue acts across dimensions). Dialogue participants spend a lot of time managing the communication. Dialogue

[30] Detailed information about the MapTask project can be found at `http://www.hcrc.ed.ac.uk/maptask/`

[31] For more information about the TRAINS corpus please visit `http://www.cs.rochester.edu/research/speech/trains.html`

control acts have a variety of functions in making communication smooth and successful, and are largely responsible for the naturalness and fluentness of spontaneous dialogue. Three major clusters of dialogue control acts are those concerned with Feedback, Interaction Management and Social Obligation Management. Feedback acts provide information either about the speaker's processing of previous utterance(s) (*Auto-feedback*) or about speaker's opinions about the addressee's processing, or solicit information about that (*Allo-feedback*). Social Obligation Management acts are concerned with social conventions and constraints. Interaction Management acts are concerned with difficulties in the speaker's contributions (*Own Communication Management*), the speaker's assistance or correction of the addressee (*Partner Communication Management*), the speaker's need for time (*Time Management*), maintaining contact (*Contact Management*), allocation of speaker role (*Turn Management*), and future structure of the dialogue (*Dialogue Structuring*).

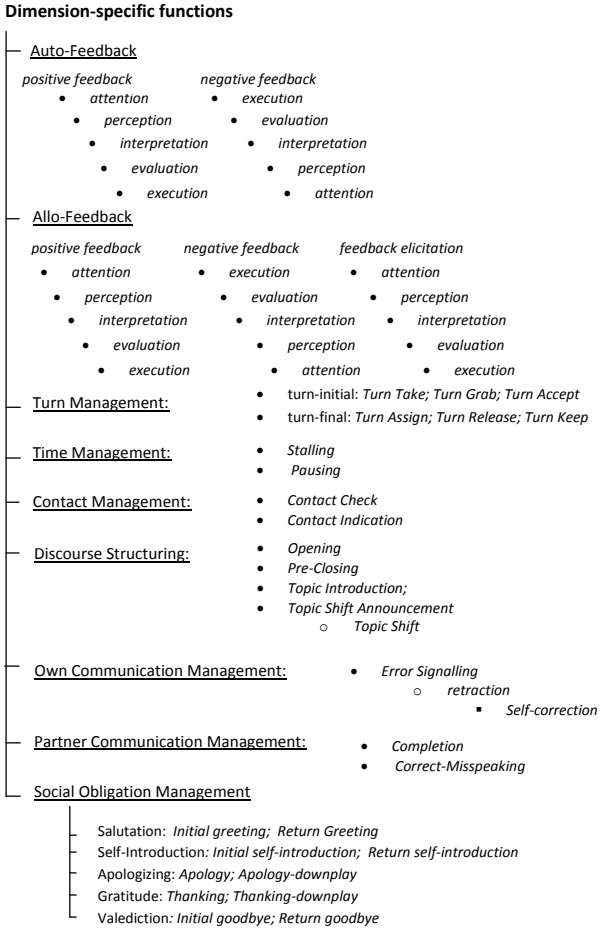


Figure 4.6: DIT^{++} dimension-specific functions.

Dimensions classify dialogue acts. What is usually called a 'dialogue act taxonomy' is

in fact a taxonomy of the *communicative functions* of dialogue acts. Dialogue act annotation is most often understood as the assignment of communicative function tags to segments of dialogue.[32] Some acts address one particular dimension. For example, a Turn Taking act is concerned with the allocation of the speaker role, and an Understanding act is concerned with the understanding of the previous utterance. Being specific for a particular dimension, these functions are called *dimension-specific*. Figure 4.6 shows the DIT^{++} taxonomy of dimension-specific functions.

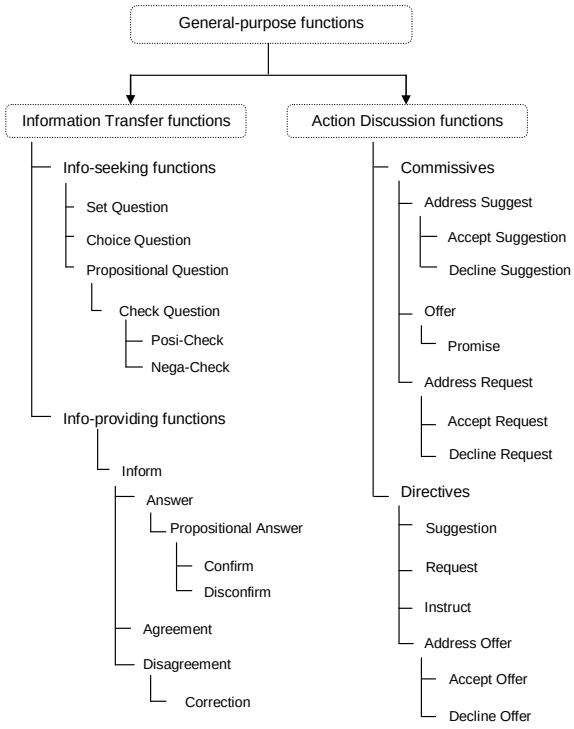


Figure 4.7: DIT^{++} general-purpose functions.

Other functions are not specifically related to any dimension, e.g. one can ask a question about any type of semantic content, provide an answer about any type of content, or request the performance of any type of action (such as '*Please close the door*' or '*Could you please repeat that*'). Question, Answer, Request, Offer, Inform, and many other functions have this property that they can be applied to a wide range of semantic content types. Given a set of dimensions, the dialogue act that results from applying the function to a particular content can be classified depending on the type of its content. Because they can be used to address any dimension, these communicative functions are called general-purpose functions. General-purpose functions are of two types: *information transfer* functions and *action discussion* functions. Information transfer functions are used to obtain (*information-seeking* functions) or to provide information

[32] When using DAMSL for dialogue annotation, one should officially assign both the top-level dimension tag (Task, Task Management, Communication Management) and the communicative function tags, but in practice only communicative function tags are applied.

(*information providing* functions). Action discussion functions have a semantic content consisting of an action, and possibly also a description of a manner or frequency of performing the action and are concerned either with the speakers commitment to perform a certain action (commissives) or his wish that the addressee performs an action (directives). Figure 4.7 shows the DIT^{++} taxonomy of general-purpose functions.
Release 4 (2008) of the DIT^{++} taxonomy has inspired the ISO act annotation standard ISO 24617-2; the specification of that standard has been developed together with the specification of Release 5 of the DIT^{++} taxonomy; formally, the latter is a superset of the former.

4.6 Conclusions

In this chapter we discussed different approaches to dialogue act annotation. The first important conclusion to draw is that multidimensional dialogue act annotations schemes do not only better capture fine-grained theoretical and empirical distinctions resulting in better coverage of dialogue phenomena, but also, contrary to what is often believed, can be applied reliably by annotators (even more reliably than one-dimensional schemes). From various annotation experiments that we performed it has been concluded that the usability and reliability of an annotation scheme is not so much a matter of the simplicity of the tag set, but rather of its conceptual clarity, with precise communicative function definitions and clear annotation guidelines (see e.g. Geertzen et al. (2008), Geertzen (2009) and Bunt et al. (2007)).

A second important conclusion concerns dialogue segmentation. The multidimensional approach to dialogue act segmentation solves various notorious problems concerning the segmentation of dialogue caused by disfluent speech, overlapping and simultaneous talk, and discontinuity of the segments that are relevant for analysis. We showed in Geertzen et al. (2007) and in Petukhova and Bunt (2011) that the multidimensional approach to segmentation results in a more accurate analysis expressed in higher scores for automatic dialogue act classification (see also Chapter 7).

The basic concepts for dialogue act annotation can be summarised in a metamodel;[33] Figure 4.8 shows the ISO metamodel for dialogue act annotation (Bunt et al., 2010).

According to its definition, a dialogue act has at least two participants: (1) an agent who produces the dialogue act, usually called the **speaker** or **sender**; and (2) a participant to whom he is speaking and whose information state he wants to influence, called the **addressee**. There may be multiple addressees. Besides sender and addressee(s), there may be various kinds of participants who witness a dialogue without actively participating. Clark (1996) distinguishes between side-participants, bystanders, and overhearers, depending on the role that they play in the communicative situation.

Dialogue acts do not occur in isolation. Some dialogue acts are semantically or pragmatically related to previous dialogue acts through **functional dependence relations** and **rhetorical relations**. Feedback acts refer to stretches of dialogue behaviour rather than to its interpretation, and are related to previous dialogue through **feedback dependence relations**.

ISO standard 24617-2 includes the definition of the Dialogue Act Markup Language DiAML. This language has a formal model-theoretic semantics associated with its abstract syntax, which rests on the assignment of information-state update schemes to communicative functions, which can be instantiated with a given semantic content (see Bunt, 2011).

[33] The term 'metamodel' is often used to describe a very general model that tries to capture the most basic notions underlying several alternative models, see e.g. Bunt and Romary (2004).

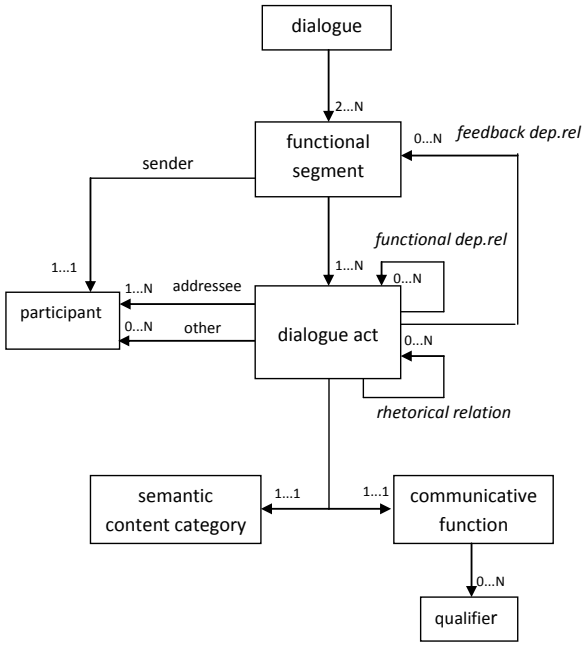


Figure 4.8: ISO 24617-2 metamodel for dialogue act annotation.

A concrete example of the use of DiAML in (44)[34]. P2's utterance is segmented into two overlapping functional segments: one in the Auto-Feedback dimension and one in the Task dimension, with value 'answer' qualified as 'uncertain'. Annotations may be attached directly to primary data like stretches of speech, defined by temporal begin and end points, but more often they will be attached to structures at other levels of analysis, such as the output of a tokenizer. TEI-ISO standard ISO 24610-1 is followed for attaching information to digital documents. In the example, the dialogue participants are assumed to be identified in the metadata of the primary data as "p1" and "p2", and their utterances are segmented multidimensionally into the functional segments "fs1", "fs2.1", and "fs2.2".

(44) a.
P1: *What time the next train to Utrecht leaves?*
P2: *The next train to Utrecht leaves I think at 8:32.*
fs2.1 The next train to Utrecht [positiveAutoFeedback]
fs2.2 The next train to Utrecht leaves I think at 8:32. [answer, uncertain]

[34]From the OVIS dialogues.

b.

```
<diaml xmlns:"http://www.iso.org/diaml/">
<dialogueAct xml:id="da1" sender="#p1" addressee="#p2"
  target="#fs1" communicativeFunction="setQuestion"
  dimension="task"/>
<dialogueAct xml:id="da2" sender="#p2" addressee="#p1"
  target="#fs2.1" communicativeFunction="autoPositive"
  dimension="autoFeedback" feedbackDependence="#fs1"/>
<dialogueAct xml:id="da3" sender="#p2" addressee="#p1"
  target="#fs2.2" communicativeFunction="answer"
  qualifier="uncertain" dimension="task"
  functionalDependence="#da1"/>
</diaml>
```

For the DIT^{++} taxonomy being a superset of the ISO inventory of concepts DiAML can be used to represent DIT^{++} more specific feedback function, that are not defined in the ISO scheme, as XML values.

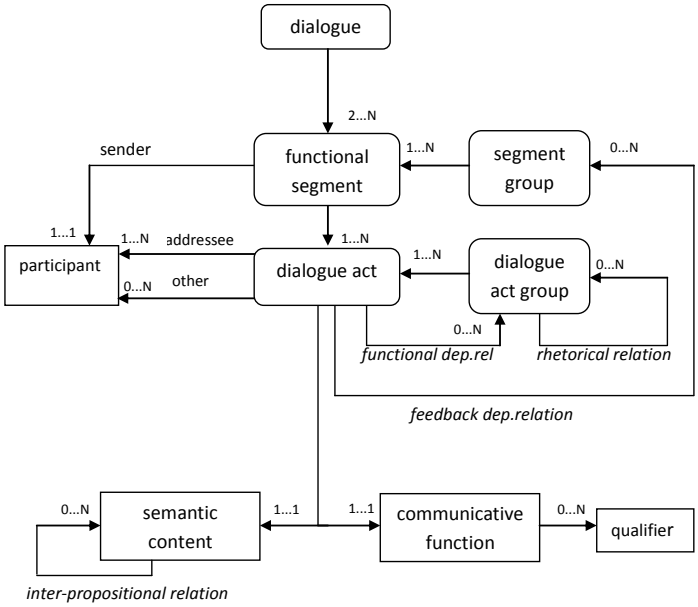


Figure 4.9: ISO 24617-2 metamodel extended with relations between dialogue units.

Our corpus study reported in Section 4.3 that considers various dialogue units and the nature of their relations, showed that functional dependence relations may occur not only between individual dialogue acts, but also between a dialogue act and a dialogue act group. Feedback dependence relations may occur not only between a dialogue act and individual functional segment, but also between a dialogue act and a group of functional segments. Rhetorical relations may occur between individual dialogue acts, between a dialogue act and a dialogue act group, but also between groups of dialogue acts. In studying the occurrence of discourse relations in

dialogue, we have observed two types of rhetorical relations: **rhetorical** relations between dialogue acts or between their semantic contents (**inter-propositional rhetorical relations**). The metamodel in Figure 4.9 has been designed as an extension of the ISO 24617-2 metamodel, containing the various kinds of units in dialogue and the possible relations between them.

DiAML can be used to represent annotations of the various kinds of relations, as illustrated in (46) for the dialogue fragment given in (45). The dialogue participants are identified as "p1" and "p2", and their utterances correspond to the functional segments "fs1", "fs2", "fs3", "fs4", "fs5", "fs6" and "fs7"; and segments "fs1", "fs2", "fs3", and "fs4" are grouped into the segment group "fsg1".

(45) P1: *We're gonna be selling this remote control for twenty five euro* [fsg1:fs1]
P1: *And we're aiming to make fifty million euro* [fsg1:fs2]
P1: *So we're gonna be selling this on an international scale* [fsg1:fs3]
P1: *And we don't want it to cost more than twelve fifty euros* [fsg1:fs4]
P2: *Okay* [fs5]
P1: *So fifty percent of the selling price*[fs6]
P2: *Can we go over that again* [fs7]

(46)
```
<diaml xmlns:"http://www.iso.org/diaml/">
<dialogueActGroup xml:id="dag1">
<dialogueActGroup xml:id="dag2">
<dialogueAct xml:id="da1" sender="#p1" addressee="#p2"
  target="#fs1" communicativeFunction="inform"
  dimension="task"/>
<dialogueAct xml:id="da2" sender="#p1" addressee="#p2"
  target="#fs2" communicativeFunction="inform"
  dimension="task" />
<rhetoricalLink dact="#da2" rhetoRelatum="#da1"
  rhetoRel="narration"/>
</dialogueActGroup>
<dialogueAct xml:id="da3" sender="#p1" addressee="#p2"
  target="#fs3" communicativeFunction="inform"
  dimension="task" />
<rhetoricalLink dact="#da3" rhetoRelatum="#dag2"
  rhetoRel="elaborate"/>
<dialogueAct xml:id="da4" sender="#p1" addressee="#p2"
  target="#fs4" communicativeFunction="inform"
  dimension="task" />
<rhetoricalLink dact="#da4" rhetoRelatum="#da3"
  rhetoRel="narration"/>
</dialogueActGroup>
<dialogueAct xml:id="da5" sender="#p2" addressee="#p1"
  target="#fs5" communicativeFunction="positiveOverall"
  dimension="autoFeedback" feedbackDependence="#fsg1"/>
<dialogueAct xml:id="da6" sender="#p1" addressee="#p2"
  target="#fs6" communicativeFunction="inform"
  dimension="task" feedbackDependence="#fs5"/>
<rhetoricalLink dact="#da6" rhetoRelatum="#dag1"
  rhetoRel="conclude"/>
```

```
<dialogueAct xml:id="da7" sender="#p2" addressee="#p1"
  target="#fs7" communicativeFunction="request"
  qualifier = "conditional" dimension="discourseStructuring"/>
<dialogueAct xml:id="da8" sender="#p2" addressee="#p1"
  target="#fs7" communicativeFunction="negativeEvaluation"
  dimension="autoFeedback" feedbackDependence ="#fsg1"/>
</diaml>
```

This example illustrates the use of DiAML for annotating relations of different type between different types of units: feedback dependence relations between a dialogue act and a functional segment (da6 and fs5); rhetorical relation between individual dialogue acts (da2 and da1; da4 and da3) and between a dialogue act and a group of dialogue acts (da3 and dag2; da6 and dag1).

Chapter 5

Forms of multifunctionality

This chapter presents an empirical account and analytical examination of the forms of multifunctionality that are found in dialogue units of various types.

As we have seen in Chapter 3, different aspects of communication that speakers may address simultaneously are identified in the ten dimensions of the DIT^{++} taxonomy. With each functional segment, several dialogue acts can be performed, belonging to different dimensions. A good understanding of the nature of the relations among the various communicative functions that a dialogue segment may have is essential for defining a computational update semantics for dialogue contributions.

Section 5.1 summarizes the four semantically different types of multifunctionality that a functional segment can have (Bunt, 2009b; 2011): independent, entailed, implicated, and default. Section 5.2 describes a corpus analysis of the occurrence of these types of multifunctionality a single functional segments; in overlapping functional segments; and in sequences of functional segments within a turn unit. Section 5.3 draws conclusions.

5.1 Semantic types of multifunctionality

5.1.1 Independent multifunctionality

A dialogue segment may have multiple functions by virtue of its observable features; this is called *independent* multifunctionality. Features include wording, prosodic and acoustic features, and accompanying nonverbal signals. For example, 'yes' and 'okay', said with an intonation that first falls and subsequently rises, express positive feedback and gives the turn back to the previous speaker. Semantically, the interpretation of a segment which displays independent multifunctionality comes down to two (or more) independent update operations on different dimensions of an addressee's information state, one for each communicative function.

The DIT^{++} tag set has been designed in such a way that two communicative functions which can be applied in the same dimension either (1) are *mutually exclusive*, or (2) one *entails*

The theoretical part of this chapter reported in Section 5.1 has been inspired by the work of Harry Bunt (see Bunt, 2010) to which I have contributed in discussions. I performed empirical analysis of multifunctionality reported in Section 5.2 which is based on Petukhova et al. (2010a), for which I did most of the writing.

the other. Consider, for example, the Time Management dimension. The speaker may need some time in order to gather his thoughts and signal a minor delay to the addressee (Stalling), or he may need to suspend the dialogue for some reason, and intend to resume after a prolonged delay (Pausing). Evidently, stalling and pausing acts are mutually exclusive: they cannot both apply to one and the same segment. Mutual exclusion is defined as follows (see Keizer et al., 2011):

(47) Two dialogue acts $A1$ and $A2$ *mutually exclude* each other iff the application of both the updates that would be caused by $A1$ and by $A2$ would result in an inconsistent context model, i.e. a state in which some proposition P is true as well as its negation.

Two functions F_1 and F_2 applied to the same semantic content p results in two logically inconsistent updates if the a set of updates of F_1 which results in updated context C_1 and the a set of updates of F_2 which results in context C_2 and from those a proposition q ($C_1 \vdash q$) and its negation $\neg q$ ($C_3 \vdash \neg q$) can be derived. This is the case when we deal with alternative end-nodes in the tag set hierarchy. For example, one cannot express a Confirm and a Disconfirm function concerning the same semantic content in one functional segment: for Confirm holds that $\mathit{believes}(S, p)$ and $\mathit{wants}(S, \mathit{believes}(A, p))$ where S stands for Speaker and A for Addressee; for Disconfirm holds $\mathit{believes}(S, \neg p)$ and $\mathit{wants}(S, \mathit{believes}(A, \neg p))$. So the updates caused by a Confirm result in a context from which can be derived that S believes that p, and the updates caused by a Disconfirm results in context from which can be derived that it is not the case that S believes that p.

5.1.2 Entailment relations between communicative functions

In the case of an entailment relation, a functional segment has a communicative function, characterised by a set of preconditions which logically imply those of a dialogue act with the same semantic content and with the entailed communicative function. Bunt (2010) defines the entailment relation between two communicative functions as follows:

(48) a. A dialogue act $A1$ *entails* a dialogue act $A2$ iff for any context model M, the update effects $||A1||_M$ on M that would be caused by $A1$ have the update effects $||A2||_M$ that would be caused by $A2$ as logical consequences.

b. A communicative function $F1$ *entails* a communicative function $F2$ iff for any semantic content a dialogue act with communicative function $F1$ entails the dialogue act with communicative function $F2$ and the same semantic content.

A particular kind of entailment relations occurs between dialogue acts within the same dimension which have the same semantic content but communicative functions that differ in their level of specificity, more specific dialogue acts entailing less specific ones. For example, Agreement and Disagreement entail Inform, and Confirm and Disconfirm entail Propositional Answer. This type of within-dimension entailment relation has also been called *functional subsumption* (Bunt, 1994).

A communicative function in one dimension may also entail a function in another dimension. Such an entailment relation occurs for example between responsive acts in non-feedback dimensions on the one hand and feedback acts on the other. Dialogue acts which respond to a dialogue act of another participant (such as accepting or rejecting an offer, suggestion, invitation, or request, answering a question, responding to a greeting, or accepting an apology) imply that the speaker believes to have understood the dialogue act sufficiently well to respond to it, and hence entail positive feedback relating to its functional antecedent.

5.1.3 Implicated communicative functions

A functional segment may also have multiple communicative functions due to the occurrence of conversational implicatures. Implicature relations between dialogue acts and between communicative functions are defined as follows:

(49) a. A dialogue act A1 implicates a dialogue act A2 iff for any context model M, the update effects $||A1||_M$ of A1 on M have the update effects $||A2||_M$ of A2 as conversational implicatures.
b. A communicative function F1 implicates a communicative function F2 iff for any semantic content a dialogue act with communicative function F1 implicates the dialogue act with communicative function F2 and the same semantic content.

For example, positive feedback is implicated by shifting to a new topic, related to the previous one; more generally, by any relevant continuation of the dialogue. Negative feedback, by contrast, is implicated by shifting to an unrelated topic; more generally, by any irrelevant continuation of the dialogue. Like all conversational implicatures, this phenomenon is context-dependent, and implicated functions are intended to be recognised. Implicated functions correspond semantically to an additional context update operation.

5.1.4 Entailed and implicated feedback functions

The levels of processing which are distinguished in DIT^{++} in relation to Auto- and Allo-Feedback have logical relations that turn up as implications between feedback acts at different levels:

(50) *attention* $<$ *perception* $<$ *understanding* $<$ *evaluation* $<$ *execution*

The implication relations between feedback at different levels are either entailments or implicatures. In the case of positive feedback, an act at level L_i *entails* positive feedback at all levels L_j where $i > j$; positive feedback at execution level therefore entails positive feedback at all other levels. Positive feedback at level L_i may *implicate* negative feedback at all levels L_j where $i < j$. For instance, when two people are talking to each other, a signal of successful perception normally implicates negative understanding. This is, however, not a logical necessity, but rather a context-dependent pragmatic matter, e.g. in human-computer dialogue the system's positive feedback at the level of perception normally does not carry this implicature. For negative feedback the entailment and implicature relations work in the opposite direction.

5.1.5 Implicit turn management functions

In addition to independent, entailed and implicated multifunctionality, there are cases which do not easily fit in any of these categories, e.g. implicit turn management functions. For example, whenever one performs a turn-initial Answer act in response to another participant's Question act, it seems inevitable that a Turn Accept act must be performed as well, even if not explicitly. Such cases are not simply entailments, since e.g. a Turn Accept would obviously not be performed if the participant already had the turn, and the question was asked in overlap with his ongoing turn.

On the other hand, the definition of an implicature is not satisfied either, since this implication cannot be cancelled. Turn-accepting or turn-taking acts if not signalled explicitly we call

side-effects of explicitly performed dialogue acts that initiate the turn. Similarly, if a dialogue participant already has the turn, every next word uttered, this may be taken as implying that he wants to keep the turn. If not signalled explicitly, turn-keeping acts are side-effects as well.

A turn releasing act is performed implicitly when a speaker stops speaking. It can be also signalled explicitly, e.g. prosodically by rising intonation followed by silence, and is then an independent function. A segment which elicits a response, such as those expressing a Question, Request, Offer, Suggestion but also initial Greeting and Good-Bye, has a *default* turn-giving function: it has this function unless there is evidence to the contrary, e.g. the speaker continues speaking.

5.2 Observed multifunctionality in dialogue units

To examine the forms of multifunctionality that occur in natural dialogue we performed a corpus analysis, using human-human multi-party interactions (AMI-meetings). We consider the occurrence of combinations of communicative functions (1) in a single functional segment; (2) in overlapping functional segments; and (3) in segment sequences within a single turn unit.

5.2.1 Multifunctionality in single functional segments

When a functional segment has several communicative functions we speak of *simultaneous multifunctionality*, following Allwood (1992). For example:[1]

(51) B1: Any of you anything to add to that at all?
A1: No
D1: I'll add it later in my presentation

In utterance B1 the speaker's intention is to elicit feedback, and the utterance also has an explicitly expressed ('any of you' plus intonation and ceasing to speak) turn releasing function. Semantically, this is a case of independent multifunctionality; the two functions, belonging to different dimensions, are both expressed by observable segment features. In utterance A1 the speaker provides an answer to B1. The speaker in utterance D1 gives no answer to B1; instead he indicates that he will provide the requested information later (a negative Auto-Feedback act at the level of execution).

A segment may also have one or more functions by virtue of its observable features and one or more functions by implication. For example:[2]

(52) B1: Just to wrap up the meeting
D1: Can we just go over the functionality again?

Utterance D1 is a request to shift the topic back to what was discussed before, i.e. a Request in the Discourse Structuring dimension. By implication the utterance has the function of negative feedback to B1, disagreeing to close the dialogue as announced in B1.

Table 5.1 gives an overview of the co-occurrences of communicative functions across dimensions for functional segments as observed in features of the behaviour (independent multifunctionality), and when entailed or implicated functions occur. We did not consider default

[1] From the AMI meeting corpus - ES2002b.
[2] From the AMI meeting corpus - ES2002b.

and side-effect communicative functions in the Turn Management dimension, since these *always* co-occur with other functions.

It can be observed that independent functions within the same dimension never co-occur. Implied functions within the same dimension always co-occur within Auto-Feedback and Allo-Feedback, where functions entail or implicate each other. Within the Task dimension implied functions often co-occur, since some general-purpose functions functionally subsume others.

Combinations of independent functions in the Time and Turn management dimensions often co-occur. A speaker who wants to win some time to gather his thoughts and wants to continue in the speaker role, may intend his stalling behaviour to signal the latter as well (i.e. to be interpreted as a Turn Keeping act). Segments also often have two independent functions in the Auto-Feedback and Turn Management dimensions. Auto-Feedback segments, such as 'Okay', 'Right', 'Alright', are frequently used at turn-initial positions with the intention to claim the speaker role. Many segments that are concerned with the dialogue task are also used to structure the discourse, e.g. announcements of the next topic in the dialogue also serve to move dialogue task forward. For this purpose, enumerative presentational/structural markers such as 'first', 'second', 'then', 'next', 'finally', are used, in particular in meetings.

Co-occurrence scores are of course higher when entailed and implicated functions are taken into account. An *implicated* function is for instance the positive feedback (on understanding and evaluating the preceding addressee's utterance(s)) that is implicated by an expression of thanks. The performance of a social obligation act often has the additional purpose of structuring the dialogue, e.g. greetings can be used to open the dialogue, and good-byes and thankings to close the conversation. Many segments that are used to structure the discourse, e.g. to shift the topic by introducing a new one, or by going back to the previous one, by implication have a positive and a negative auto-feedback function, respectively.

It can be also observed that Time Management is never implied by other acts. This means, that Time Management acts always expressed explicitly, e.g. verbally by using filled pauses and word lengthening. Time Management acts may have various implicatures (see also Clark and Fox Tree, 2002). For example, an extensive amount of stallings accompanied by relatively long pauses may be intended to elicit support for completing an utterance (Own Communication Management), to invite the addressees to speak (Turn Management), to obtain the addressee's attention (Allo-Feedback), or to indicate contact (Contact Management).

Table 5.1: Co-occurrences of communicative functions across dimensions in one functional segment, expressed in relative frequency in %, when implied functions (implicated and entailed) are excluded and included. (Read as follows: percentage of segments having a communicative function in the dimension corresponding to the column, which also have a function in the dimension corresponding to the row.)

	form	Task	Auto-F.	Allo-F.	Turn M.	Time M.	DS	Contact M.	OCM	PCM	SOM
Task	independent	0.0	1.1	0.0	2.2	2.4	19.6	0.0	69.9	0.1	0.0
	implied	49.8	47.9	24.9	0.0	0.0	31.5	0.4	0.0	0.0	0.7
Auto-F.	independent	0.7	0.0	0.0	11.0	0.6	1.9	11.1	0.8	0.0	0.0
	implied	38.9	100.0	0.0	0.0	0.0	11.2	20.2	11.7	65.0	80.0
Allo-F.	independent	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
	implied	24.9	0.0	100.0	0.0	35.7	2.1	1.2	7.9	0.7	10.3
Turn M.	independent	3.4	26.9	6.7	0.0	28.6	12.4	7.4	4.8	18.2	6.7
	implied	0.0	0.0	0.0	0.0	11.4	0.0	0.0	0.0	0.0	0.0
Time M.	independent	28.2	11.3	7.8	44.9	0.0	4.7	0.0	83.23	0.5	0.0
	implied	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DS	independent	0.1	0.4	0.0	0.3	0.0	0.0	0.9	3.7	0.0	6.7
	implied	3.2	58.3	29.1	0.0	0.5	4.6	25.0	0.0	0.0	32.5
Contact M.	independent	1.7	0.3	0.0	3.6	0.5	3.7	0.0	0.3	0.0	1.3
	implied	2.4	97.1	1.6	0.0	4.9	2.4	0.0	0.0	0.0	7.6
OCM	independent	1.2	0.4	0.0	2.8	0.5	0.0	0.0	0.0	0.9	6.7
	implied	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PCM	independent	0.1	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0
	implied	0.0	15.0	11.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SOM	independent	0.0	0.0	0.0	0.2	0.0	0.0	2.7	0.3	0.0	0.0
	implied	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5.2.2 Multifunctionality in overlapping segments

Participants clearly do not limit their dialogue contributions to functional segments; their goal is to produce coherent utterances. An utterance may contain overlapping (or embedded, as a special case of overlapping) functional segments with different communicative functions, a larger segment having communicative function F1 containing a smaller segment that has communicative function F2. For example:[3]

(53) B1: I think we're aiming for the under sixty five
D1: **Under sixty five** is a good constraint

The functional segment formed by utterance D1 has the function of positive feedback about utterance B1 at the level of evaluation, whereas the part marked in bold is an explicit feedback signal at the level of perception. Such a co-occurrence is possible because higher levels of positive feedback entail lower levels of positive feedback (in order to evaluate an utterance one needs to pay attention to, perceive and understand what has been said).

The most important sources of overlapping multifunctionality are feedback functions, expressed explicitly by means of certain utterance features. For instance, answers often overlap with explicitly expressed positive feedback, e.g. when the speaker repeats (positive perception) or paraphrases the partner's previous utterance (positive interpretation), see Table 5.2 for co-occurrences with the Auto-Feedback dimension. For example:[4]

(54) D1: Which is the clunky one on the left or on the right?
C1: **The clunky one** is the one on the right

The speaker of C1 could have said just *'on the right'*, which would be a perfectly acceptable answer to the question D1. Instead, he repeats part of the question and thereby signals that his perception was successful. In the same way, Accept and Reject Offer, Suggestion and Request, but in fact any responsive act, which entails positive auto-feedback, may overlap with such segments.

Another source of overlapping multifunctionality is formed by conversational implicatures. It is often possible to add explicitly what is implicated without being redundant. For example, positive feedback implicated by shifting to a new topic, related to the previous one, may be expressed explicitly and happens very often by means of discourse markers, such as 'and then', 'okay then', 'next', etc. (see Petukhova and Bunt, 2009b). More generally, any relevant continuation of the dialogue implicates positive feedback, and this may also be expressed explicitly by repeating or paraphrasing part of a previous utterance, and using discourse markers like 'then'. For example:[5]

(55) D1: This idea focuses on the twenty five age group
B1: Are we aiming at **a fairly young market then**?

[3] From the AMI meeting corpus - ES2008b.
[4] From the AMI meeting corpus - ES2002c.
[5] From the AMI meeting corpus - ES2008b.

Table 5.2: Co-occurrences of communicative functions across dimensions in overlapping segments, expressed in relative frequency in %.(Read as follows: percentage of segments having a communicative functions in the dimension corresponding to the column, which also has a function in the dimension corresponding to the row.)

	Task	Auto-F.	Allo-F.	Turn M.	Time M.	Contact M.	DS	OCM	PCM	SOM
Task	0.0	40.8	23.4	42.4	38.2	0.0	28.2	65.4	0.0	18.2
Auto-F.	10.5	6.7	16.9	16.9	19.1	18.8	19.1	14.2	0.8	9.5
Allo-F.	1.5	4.2	1.3	4.3	12.1	18.8	12.1	5.4	16.2	9.1
TurnM.	14.1	31.4	45.9	0.0	14.6	25.0	14.6	0.0	0.0	4.9
TimeM.	2.9	7.7	20.2	12.8	0.0	0.0	0.8	3.4	16.1	3.2
ContactM.	0.3	0.2	1.8	0.1	0.0	0.0	5.6	0.0	0.0	2.9
DS	2.1	6.9	11.4	0.2	3.9	7.5	0.0	5.6	0.0	8.2
OCM	4.6	3.8	5.8	4.4	2.3	0.0	2.2	0.0	0.0	1.6
PCM	0.0	0.9	0.9	1.2	0.7	0.0	0.7	0.0	0.0	0.0
SOM	0.0	0.1	1.3	2.1	0.3	3.3	0.3	0.2	0.0	0.0

Table 5.2 gives an overview of the co-occurrences of communicative functions across dimensions for overlapping functional segments. Auto-Feedback segments are often embedded in segments with functions in other dimensions, as illustrated above. Similarly, Turn Management acts may be embedded into larger segments that form acts in other dimensions. Task acts often embed smaller segments by which a variety of dialogue acts are performed. For example, this is often the case with self-corrections or retractions (Own Communication Management), when the speaker recognizes that he made a mistake. In this case the speaker normally stops the flow of the speech and signals (e.g. using editing expressions) that there is trouble and that the repair follows.

5.2.3 Multifunctionality in segment sequences within a turn unit

Functional segments may be *discontinuous*, where one segment breaks off another one. For example, the speaker in (56) interrupts his Inform with a Set -Question:[6]

(56) Twenty five euros for a remote... **how much is that locally in pounds?** is too much to buy a new one

Functional segments following each other within a turn give rise to what Allwood (1992) calls *sequential multifunctionality* at turn level. We analysed sequences of a length of 2 functional segments for the most frequently occurring patterns of communicative function combinations (see Table 5.3).

The co-occurrence scores for Turn Management, Task and Auto-Feedback with other dimensions are relatively high. Task-related functional segments are frequently preceded or followed by Turn Management or Auto-Feedback segments, or segments that have functions in these two dimensions simultaneously. For instance, a frequent pattern for constructing a turn is first performing a turn-initial act (Turn Take, Accept or Grab) combined with or followed by an Auto-Feedback act and one or more segments in another dimension, and closing up the turn with a turn-final act. This pattern occurs in 49.9% of all turns. For example:[7]

(57) B1: Well (Neg.Auto-Feedback Evaluation + Turn Take)
B2: Twenty five euros is about eighteen pounds, isn't it? (Auto-Feedback Check Question + Turn Release)
D1: Um (Turn Take + Stalling)
D2: Yep (Allo-Feedback Confirm)

Performing a sequence of dialogue acts within one turn unit, dialogue participants generally order the corresponding segments in a coherent fashion. To first reject a request and subsequently accept it would be strange, for example unless the first act is performed by mistake or the speaker changes his mind, and withdraws the acceptance.

We often observed sequences where the speaker performs a certain act and subsequently tries to justify or elaborate it, or explains what he just said. For example:[8]

(58) A1: It ties you on in terms of the technologies
A2: **Like for example** voice recognition
A3: **Because** you need to power a microphone
A4: **So** thats one constraint there

[6] From the AMI meeting corpus - ES2002a.
[7] From the AMI meeting corpus - ES2008a.
[8] From the AMI meeting corpus - ES2002c.

Table 5.3: Co-occurrences of communicative functions across dimensions in a sequence of two functional segments in one turn, expressed in relative frequency in %. (Read as follows: percentage of segments having a communicative functions in the dimension corresponding to the column, which also has a function in the dimension corresponding to the row.)

	Task	Auto-F.	Allo-F.	Turn M.	Time M.	DS	Contact M.	OCM	PCM	SOM
Task	26.5	36.5	33.3	33.5	42.4	0.0	15.4	21.6	20.0	46.7
Auto-F.	15.9	24.8	9.9	16.7	17.2	33.3	19.2	8.0	30.0	13.3
Allo-F.	0.4	1.1	6.6	0.6	0.6	0.0	0.0	0.5	0.0	0.0
TurnM.	59.7	38.1	36.7	53.0	44.2	15.3	61.5	69.9	50.0	33.3
TimeM.	27.9	20.4	20.0	30.9	18.8	0.0	15.4	55.4	0.0	26.7
ContactM.	0.0	0.1	0.0	0.1	0.0	34.2	0.0	0.0	0.0	54.6
DS	0.5	1.2	0.0	0.6	0.6	15.0	7.6	0.5	0.0	0.0
OCM	9.9	8.0	6.7	11.3	13.9	0.0	7.7	9.5	0.0	0.0
PCM	0.4	0.42	0.0	0.1	0.1	0.0	0.0	0.3	0.0	0.0
SOM	0.2	0.6	0.0	0.3	0.1	33.3	0.0	0.5	0.0	6.7

In example (58) discourse markers are used by the speaker to indicate the steps in a sequence of arguments: he makes a statement (Inform); then provides an example for this statement (Inform Exemplify); justifies his choice (Inform Justification); and draws a conclusion (Inform Conclude).

5.3 Conclusions

In this chapter we have described different forms of multifunctionality that occur in natural dialogue and the relations between the communicative functions of multifunctional dialogue units. One of the main conclusions is that any adequate account of the meaning of dialogue utterances has to take their multifunctionality into consideration, since many functional segments display some form of multifunctionality.

Functional segments display both independent multifunctionality, having two or more functions in different dimensions due to their features, as well as implied multifunctionality. The latter occurs if communicative functions have certain entailment relations, conversational implicatures, default functions, or side-effects.

Another conclusion is that there are certain co-occurrence patterns of dialogue acts addressing different dimensions within a single functional segment, in overlapping or embedded segments, and sequences of functional segments within a single turn unit. This good news for computational dialogue modelling and for automatic dialogue act recognition. For the former it can facilitate the effective computation of dialogue act combinations and the specification of combinatorial constraints. For the latter it can help to reduce the search space.

Semantically, the interpretation of a segment which displays independent multifunctionality comes down to two (or more) independent update operations on different dimensions of an addressee's information state, one for each communicative function. The update operations of entailed functions within one dimension are subsumed by those of the entailing function. Entailment relations between non-feedback acts and auto- and allo-feedback acts, by contrast, correspond to additional context updates.

Implicated functions correspond semantically to additional context updates, since they are logically independent of the functions that implicate them.

The default functions of stopping speaking as constituting a Turn Release and the side-effect functions of starting speaking as constituting a Turn Take or a Turn Accept act and continuing speaking as constituting a Turn Keeping act are additional context updates, but given their consistent co-occurrence with other functions such updates can be added automatically for every segment, or token if incremental dialogue act interpretation is wanted (see Chapter 7.

This study contributes to the definition of the interpretation and generation of dialogue utterances. However, corpus observations only are not sufficient. Additional analytical examinations of dialogue act preconditions, entailments and implicature relations will be provided in Chapter 8.

Chapter 6

Multimodal forms of interaction management

In this chapter we analyse how dialogue participants express the intended functions of their dialogue contributions concerned with interaction management, using multiple modalities such as speech and body movements. The main objective is to identify those features of the physical realisation of interaction management utterances in context that enable their recognition. We report results of explorative studies, observations from annotated data, and perceptual experiments.

Introduction

Dialogue participants use all the modalities available for them in dialogue. In telephone conversations, where only the sound modalities are available, participants use linguistic and vocal devices to express their intentions. Such devices include particular lexical and syntactic devices, intonation and loudness, and sounds like laughing, sighing, and coughing. In face-to-face interaction nonverbal communication is as important as verbal communication; it includes the use gesture, facial expression, gaze direction, head orientation, posture, and touch.

Verbal actions have been studied extensively, and relatively much effort has been spent on their computational modelling (see e.g. Reithinger, 1997 and Webb et al., 2005; Shriberg et al., 1998; Jurafsky et al., 1998a; Fernandez and Picard, 2002; Stolcke et al., 2000). Relatively little research has been devoted to the computational modelling of nonverbal communication. This is partly because the facilities for recording and analysing human movements, needed for the study of bodily communication, have become available only recently. Another reason, mentioned by Allwood (2002), is that monologue has traditionally been emphasised over dialogue, and speech over body. A large part of the dialogue acts which do not directly relate to the dialogue task, but that serve to manage the interaction, are expressed nonverbally, however.

This chapter is based on Petukhova and Bunt (2009c); Petukhova and Bunt (2009b); Petukhova and Bunt (2009a); and Petukhova and Bunt (2010c). These conference papers are largely written by me, with continuous support from the co-author. Experiments described in this chapter are designed and carried out by me with support of my Master students: Frederike Groothoff, Véronique Verhagen and Karin Fikkers.

While it is usually taken for granted that nonverbal activity is an essential ingredient of interaction, it has so far resisted an integrated formal account. For many applications the extensive analysis and modelling of properties of multimodal behaviour would be a very useful asset, e.g. for embodied conversational agents and communicating robots.

In order to be able to integrate nonverbal communication in a computational dialogue model, we explored the semantic and pragmatic information that is available in nonverbal modalities and determined and described the role of nonverbal signals in dialogue, focusing on the following aspects: (i) what type of information is transmitted by different modalities; (ii) what are communicative functions of the multimodal utterance as a whole; and (iii) in what way do different modalities interact and contribute to the communicative functions of multimodal utterances.

This chapter is organised in the following way. Section 6.1 discusses the verbal and nonverbal data that we analysed, and describes the coding of the observed low-level behavioural features relating to different modalities. Subsequently, we analyse multimodal expressions of dialogue acts that address the most frequently occurring dialogue control dimensions: feedback acts (Section 6.2), turn management acts (Section 6.3) and discourse structuring acts (Section 6.4). The role of nonverbal signs in general is discussed in Section 6.5. The main observations and conclusions of the analyses are summarized in Section 6.6.

6.1 Multimodal expression of dialogue acts

What was said in dialogue is of crucial importance, but also how it was said. For all studied corpora, the speech transcriptions were provided with corpus data and contain manually produced orthographic transcriptions, including word-level timings. Not only single words are interesting for the analysis, but also their collocation and co-occurrence patterns. For this, bi- and trigram models were constructed.

For each token prosodic properties were computed automatically using PRAAT tool for voice analysis (Boersma and Weenink, 2009). Computed prosodic properties are minimum, maximum, mean, and standard deviation of *pitch* (F0 in Hz), *energy* (RMS), *voicing* (fraction of locally unvoiced frames and number of voice breaks) and speaking rate (number of syllables per second). We examined both raw and normalized versions of these features. Speaker-normalized features were obtained by computing z-scores (z = (X-mean)/standard deviation) for the feature, where mean and standard deviation were calculated from all functional segments produced by the same speaker in the dialogues. We also used normalizations by the first speaker turn and by prior speaker turn. Additionally, for each token temporal and durational properties were considered: token *duration* and *floor-transfer offset*[1] computed in milliseconds.

Our study focuses on five forms of nonverbal expression: gaze direction, head movements, hand and arm gestures, posture shifts, and facial expressions.

Gaze shows the focus of attention of the dialogue participant. Gaze is also an important signal of liking and disliking, and of power and status. For example, if two people of different power or status meet, the low-power person looks at the other much more as he listens than as he talks, while there is no such difference for the high-power individual (Argyle, 1994). Gaze is also used to ensure contact between participants, for example, the speaker looking at an addressee signals that he is interested in his attention, wanting him to be involved. For this

[1] Difference between the time that a turn starts and the moment the previous turn ends.

purpose so-called 'mutual gaze' is used, people looking at each other for some time. Participants break 'mutual gaze' when they close the interaction.

Patterns of gaze have been found to be most strongly correlated with verbal turn-taking behaviour. Exploring gaze behaviour in informal two-party conversations, Kendon (1967) noticed that people change their gaze direction in consistent ways around utterance and turn beginnings and endings. The observed speakers tended to look away from their dialogue partner as they began a new utterance, or slightly before this. When approaching the end of the utterance, they would often look up to their listener, which frequently coincided with a listener averting his or her gaze from the speaker. Duncan (1970) and Wiemann & Knapp (1975) found steady increases of the frequencies of listener-directed gazes during speaking turns, and a drop in the duration of speaker-directed gaze during the last third of speaking turns. These changes in gaze direction have been interpreted as signals of speakers and listeners about their intentions to start or finish a turn. By directing his or her gaze before finishing an utterance a speaker can signal this end in advance, offering a partner the possibility to speak, and monitor whether the partner intends to make use of this possibility. By averting his or her gaze during this phase of the utterance the speaker can make clear that he or she is planning a next utterance.

In multi-party conversation nonverbal behaviour, in particular gaze, plays a more significant role in managing fluent turn transitions than in two-party dialogues, because of an increased uncertainty about who the next speaker will be. Using nonverbal rather than verbal actions to regulate turn transitions would interfere less with verbal communication (Vertegaal et al., 2001). Differences have been observed in the gaze behaviour of participants in multi-party conversations from that in two-party dialogue. Participants direct their gaze less frequently while speaking than while listening during two-party dialogues (Exline (1963), Kendon (1967), Argyle (1994)). The great amount of directed gaze during speaking could be the result of using directed gaze to make clear to whom one is speaking, a need that arises only in group conversations.

Head movements and head orientation are the basic forms of signalling understanding, agreement and approval, or failure. Head nods, shakes, turns, and jerks have been distinguished as actions performed by listeners to provide speakers with feedback on their message (Duncan, 1972). It has also been suggested that these head movements are responses to head movements of speakers, who may use this as a means to request feedback (McClave, 2001). Feedback functions of head movements can thus interact with turn management functions. Hadar et al.(1984) investigated whether it is likely that head movements are used for the latter purpose. They reported that the vast majority of head movements (89 out of 99) was performed by speakers rather than by listeners. Most of the speaker's head movements were located around initiations of speech after breaks between either syntactic clauses or turns. They concluded that speakers use head movements both to mark syntactic boundaries and to regulate the process of turn-taking.

Head movements are also used to indicate aspects of information structure, e.g. to mark alternatives, or contrast; or to express a cognitive state, e.g. uncertainty or hesitation. Heylen (2006) noticed that head movements may have a clear semantic value, and may mark interpersonal goals and attitudes.

Hand and arm gestures have been studied extensively, especially for their relation to the semantic content of an utterance (see e.g. Kendon, 2004; McNeill, 1992; Ekman and Fiesen (1981). Hand and arm gestures may also have interactive functions, especially, when aligned with speech in such a way that they are finished before the end of the turn. Stopping to gesticulate can be recognized by the hand dropping into a resting position, or the relaxation of

a tensed hand position. These movements can therefore serve as a signal that the turn will soon end. Since co-speech gestures can make clear that a speaker is not about to finish talking, their presence can signal a Turn Keep function (Duncan, 1972). The beginnings of gesticulations have been observed to mark turn-initial acts (Petukhova, 2005).

Posture shifts are movements or position shifts of the trunk of a participant, such as leaning forward, reclining, or turning away from the current speaker. Posture shifts occur in combination with changes in topic or mode of participation (e.g. Scheflen (1964), Condon and Osgton (1971), Erickson (1975), Hirsch (1989)). Cassell et al. (2001) found that both turn boundaries and discourse segment boundaries had an influence on the occurrence of posture shifts. Posture shifts occur more frequently, and tend to be more energetic, at discourse unit boundaries than within discourse units.[2] Also, participants were shown to be five times more likely to show posture shifts at a turn boundary than within a turn. When a participant simultaneously starts a new turn and a new discourse unit, this is marked with a posture shift ten times more often than when a participant starts a new turn within the same discourse unit. As such, posture shifts may be more related to discourse structure than to turn management.

Facial expressions are important for expressing emotional reactions and attitudes to other people, such as happiness, surprise, fear, sadness, anger and disgust or contempt (Argyle, 1994). These six basic emotions are found in all cultures. Face can also display a state of cognitive processing, e.g. disbelief, surprise, or lack of understanding.

Little research has been done on the function of facial expressions as cues for interaction management. Wiemann and Knapp (1975) suggest that for the organisation of turns smiles might be important. In general, however, possible relations between facial expressions and interaction organisation form an uncharted territory.

6.1.1 Coding visible movements

The transcriptions that we made of visible movements include gaze direction, head movements, hand and arm gestures, facial expressions (including constriction or relaxation of forehead muscles, eyebrow movements, changes in eye shape, and lips movements), and posture shifts. Low-level behavioural features were coded such as *form of movement* (e.g. head: nod, shake, jerk; hands: pointing, shoulder-shrug, etc.[3]); *direction* (up, down, left, right, backward, forward); *trajectory* (e.g. line, circle, arch); *size* (e.g. large, small, medium, extra large); *speed* (slow, medium, fast); and *repetitions* (up to 20 times). For each movement *intensity* was determined: 0 - no movement; 1 trace (noticeable movement); 2 marked (significant evidence for a movement).

The nonverbal behaviour of the dialogue participants was transcribed using video recordings for each individual participant, running them without sound to eliminate the influence of what was said.

Transcriptions were performed using the ANVIL tool[4]. The ANVIL tool allows transcriptions in multiple tiers so that for each participant we specified speech tier and several tiers for each type of movement. Moreover, ANVIL tool makes possible the multidimensional segmen-

[2]Discourse unit is a group of dialogue acts that are concerned with a particular discussion topic, sub-dialogue or sub-task. This is not the same as the notion of 'discourse unit' proposed by [Traum, 1994] for describing grounding in dialogue, which consists of an initial presentation and as many utterances as needed to make the initial utterance mutually understood. The two notions sometimes coincide, but not in general.

[3]Hand gesture transcription was performed according to Gut et al. (2003).

[4]For more information about the tool visit:
`http://www.dfki.de/~kipp/anvil`

Table 6.1: Cohen's kappa scores for each type of visible movement reached by two coders.

Type of movement	Kappa
Gaze	.8
Head movements	.79
Hand movements	.42
Facial expression	.62
Posture shifts	.78

tation of dialogue units into functional segments and their annotation (labelling) in multiple dimensions simultaneously.

Transcriptions of visible movements were made by two coders in order to assess inter-coder agreement. Inter-coder agreement was measured in terms of standard kappa. Table 6.1 presents the kappa scores.

6.2 Feedback acts

Conversation (i.e. speaking and listening) is a **bilateral process** - that is, a joint activity, and speaking and listening are not autonomous processes - conversational partners monitor their own processing of the exchanged utterances as well as the processing done by the others. "Speakers monitor not only their own actions, but also those of their addressees, taking both into account as they speak" (Clark and Krych, 2004). Given the bilateral nature of conversation, interlocutors can construct and provide feedback on both their own processing (*auto-feedback*) as and on that done by the other (*allo-feedback*).

Feedback is crucial for successful communication. Allwood et. al (1993) characterize feedback as a mechanism that speakers use to manage the flow of an interaction. Feedback can have two **polarities**: *positive*, acknowledging that communication works well, and *negative*, signalling that there is a communication problem. Sometimes also 'neutral' feedback is distinguished. For example, in DAMSL the communicative function 'Maybe' is defined for responsive acts where the speaker does not express his agreement or disagreement with the previous proposal. In DIT feedback acts are concerned with the processing of previous contributions, and processing is either successful or encounters a problem.

Feedback can be provided at different **levels** of processing the communicative behaviour of interlocutors. Allwood et al. (1993) and Clark (1996) notice that interlocutors need to establish *contact* and gain or pay *attention* to each other's behaviour, in order be involved in conversation. A speaker's behaviour needs to be *perceived* (i.e. heard, seen) or *"identified"* (Clark, 1996). Perceived behaviour should be *interpreted*, i.e. interlocutors should be able to extract the meaning of each other's behaviour. The constructed interpretation needs to be *evaluated* against one's information state: if it is consistent with the current information state it can be incorporated into that state; if it is inconsistent, this can be reported as negative feedback. The incorporation of new information, and the performance of other mental and physical actions in response to communicative behaviour is called the *execution* or *application* level of processing (Bunt, 2000).

Another aspect of feedback is its **direction**. A speaker in dialogue may provide feedback (***feedback giving***) or elicit feedback (***feedback eliciting***). With respect to one's own processing one can only give feedback (auto-feedback is always feedback giving), but with respect

to an addressee's processing a speaker can give feedback, expressing his beliefs about the addressee's success in processing a previous utterance, and he can also elicit feedback, if he wants to know for example if the addressee understood him.

Feedback can be given *explicitly* or *implicitly*. Implicit feedback is not directly detectable in the speaker's behaviour, but can be inferred. For instance, any relevant dialogue continuation as a rule is taken as a sign that the previous utterance(-s) was processed successfully. Explicit feedback can be provided either *linguistically* (e.g. through the wording of an utterance, or prosodic and acoustic form) or *non-verbally* (e.g. through gaze direction, facial expressions, head movements). In fact, feedback is mostly expressed simultaneously by vocal/verbal and gestural means (Petukhova, 2005), combining the two means of expression multimodally.

Explicit feedback utterances may be more or less expressive and elaborate. Feedback can given or elicited in an *inarticulate* way, e.g. 'What?' or 'Huh?', or in an *articulate* way, e.g. 'Do you mean this Thursday?'.

Auto-feedback is a reaction to a contribution of another interlocutor, with which it may overlap. One might expect that non-verbal feedback can easily be given simultaneously, whereas verbal feedback is more likely to be given sequentially, after the contribution is finished and the speaker is letting go of the floor. Ford and Thompson (1996) found that verbal feedback most often occurs at Complex Transition Relevant Places (CTRP); places where intonation groups coincide with pragmatic and with syntactic/interactional boundaries. Regarding the temporal position of non-verbal feedback in relation to the previous contribution, not much research has been done. We address the following questions: (1) In what way do dialogue participants signal positive auto-feedback, verbally and nonverbally? (2) What is the position and motivation of the feedback utterances in dialogue?

6.2.1 Inarticulate feedback

By inarticulate feedback we mean feedback that is expressed with minimal lexical and/or non-verbal material, e.g. utterances like 'Okay', 'mm-mhm', 'yeah' or head movements, facial expressions, or combinations of those. Such feedback acts have no or only marginal semantic content; their meaning is concentrated in their communicative function. Due to the minimalistic nature of such feedback it is sometimes difficult to figure out at what level of processing the feedback is provided: whether the speaker signals active listening, interpretation was constructed successfully and evaluated, or the information was adopted. Inarticulate feedback can be also negative.

This type of feedback also attracted a lot of attention of researchers working in the area of Embodied Conversational Agents (ECAs), i.e. in the SEMAINE project,[5] which aims to build a Sensitive Artificial Listener that provides audiovisual listener feedback in real time while the user is speaking, and takes the user's feedback into account while the agent is speaking. To assure successful communication, such artificial agents should be able to exhibit appropriate behaviour when playing the role of the listener in a conversation with a user and provide responses about their perception, attention, interest, understanding, attitude and acceptance towards what the speaker is saying.

Inarticulate feedback is the most frequently used type of feedback occurring in our dialogue data. In AMI dialogues 9.4 feedback acts are performed per minute of dialogue conversation on average. Relatively little feedback is given during the meeting opening phase (2.4); many

[5]For more information visit `www.semaine-project.eu`

Table 6.2: Distribution of expressions (relative frequency in %) of positive inarticulate auto-feedback in the analysed AMI meeting corpus

Verbal feedback		Non-verbal feedback		Verbal + non-verbal feedback	
Expression	Frequency	Expression	Frequency	Expression	Frequency
(a) okay	17.6	(b) face	3.4	a+b	4.7
(c) mm-hmm	7.4	(d) head nods	18.1	c+b	2.8
(e) (al)right	2.8	(g) b+d	8.1	a+g	17.5
(f) yeah/yep	14.4			c+d	7.8
				e+d	3.1
				f+b	2.2
				f+d	7.3
Total	24.2		29.6		46.2

feedback acts occur when important issues are discussed (13.9) and near the ending of the meeting (13.1).

Feedback expressions

Positive inarticulate auto-feedback is most of the time expressed non-verbally or by combinations of verbal expressions with non-verbal ones. Table 6.2 presents the relative frequency of all forms and expressions of inarticulate positive auto-feedback in the AMI corpus, and shows that head nods are the most frequent form.

Head nods comes in many variations, differing in speed, number of repetitions, and size or depth: fast multiple short nods, fast single short nods, slow multiple long nods and slow single long nods. We analysed the size, speed and number of repetitions of nods. The results of this study are presented in Table 6.3. Regarding the size of the nods, it was found that 63.6 percent are small, 22.7 percent are medium sized, and 13.7 percent are large nods. With respect to the speed of nodding, 12.6 percent were found to be slow, 27.6 percent medium and 59.8 percent fast. Regarding the number of repetitions, the single nod occurred most often (38.7%), followed by nods that are repeated once (22.7%) and those repeated five or more times (22.7%).

Table 6.3: Frequency (in %) of different types of nodding with respect to size, speed and number of repetitions in the analysed AMI meeting corpus

Size		Speed		Repetitions	
Short	63.6	Slow	12.6	1	38.7
Medium	22.7	Medium	27.6	2	22.7
Long	13.7	Fast	59.8	3	9.1
				4	6.8
				≥ 5	22.7

Facial expressions are complex signals, constructed out of the following main components: forehead (constricted or relaxed), changes in eye shape (smiling/narrow eyes, blinking), and lip movements (elongate and half-open). Face expresses interest, surprise, acceptance/agreement, approval, etc.

Table 6.4: Position of positive inarticulate auto-feedback in relation to main partner utterance in the analysed AMI meeting corpus (proportions in %, timing in brackets in milliseconds)

	Overlapping	Not-overlapping	Overlap with non-verbal feedback	Overlap with with next speaker	Total
Verbal f.	3.1(-565)	9.8(172)	2.5(-30)	1.2(649)	16.6
Non-verbal f.	36.8(-1220)	10.4(160)	4.3(-277)	7.4(1280)	58.9
Combination	11.0(-894)	3.7(356)	5.5(-315)	4.3(467)	24.5
Total	50.9(-893)	23.9(229)	12.3 (-207)	12.9(799)	

Negative inarticulate feedback at the level of attention is generally characterized by absence of any noticeable verbal or nonverbal activity of the dialogue participant or when the participant's focus of attention is directed to a dialogue partner other than the current speaker. The speaker, in such cases, may attract attention from his interlocutors by making pauses and looking at them, leaning to the intended addressee or making sharp hand movements. Negative feedback at the level of perception is often signalled by puzzled facial expression (curving the mouth downward, lowering the eyebrows and eyelids, dropping the jaw, constricting the forehead muscles), cupping the ear hand gesture (meaning 'I can't hear you'). Negative feedback at higher levels is signalled by head shakes (signalling opposition or inability to perform a requested action), and raising the shoulders (meaning 'I don't know' or 'Maybe'), waggles (head movements back and forth or left to right signalling uncertainty), lip-pout or compression (signalling disappointment, disbelief, disliking or disagreement), or lowering eyebrows (indicator of skepticism, disagreement or doubt).

Position and timing of feedback

Being a reaction to a contribution of another interlocutor, an auto-feedback act can either overlap with this contribution or follow it. We examined positive auto-feedback acts, since the occurrences of negative inarticulate auto-feedback are rare in our data. For all positive auto-feedback acts we examined what the relation is with the 'main' partner utterance, distinguishing between purely verbal feedback, non-verbal feedback, and verbal combined with non-verbal feedback.

Feedback utterances are frequently used around the segment boundaries of the main speaker: (1) in final boundary position in 39.4% of the cases; (2) near the start of a new segment after turn-allocation or turn continuation signals like discourse markers (e.g. 'so', 'and', 'because', 'such as', 'but'); editing expressions, restarts, or retractions, in 22.3% of all cases; (3) during turn-internal hesitation phases (36% of all cases).

Table 6.4 shows the overlapping behaviour of auto-feedback acts with respect to the 'main' partner utterance. The majority of the positive auto-feedback provided by the interlocutors is non-verbal and overlapping. If the various types of feedback are examined separately, it can be observed that non-verbal feedback mostly overlaps with speech of another interlocutor; either the current speaker's utterance or speech of the next speaker. By contrast, the verbal feedback acts of the interlocutors in most cases do not overlap. Feedback which is both verbal and non-verbal exhibits overlap approximately two-thirds of the time. It must be remarked that this kind of feedback has been examined as a whole, i.e. it has not been determined whether it is just the non-verbal part of this feedback which overlaps with the speech of another interlocutor.

One of the reasons for providing feedback might be that the interlocutor is selected by the speaker as his/her primary addressee and therefore invited to play the role of feedback giver. We tested this hypothesis by analysing the direction of gaze of the four interlocutors. Analysing gaze provides insight with regard to the question whether an interlocutor has a stronger tendency to give feedback when the speaker looks at him while talking.

For each positive feedback act we examined the direction of gaze of the current speaker. If the speaker was looking at the interlocutor who provides feedback, this is counted as 'looked at by speaker'. The speaker can, in this case, also look at multiple interlocutors while speaking and they can all provide feedback subsequently. When the speaker was not looking at the interlocutor in question, it is counted as 'not looked at by speaker'. In addition to these two categories, there is the category 'unspecified', which contains feedback acts following speaker's utterances for which the direction of gaze could not be determined.

As shown in Table 6.5, the majority of positive auto-feedback is provided in reaction to the directed speaker's gaze to the feedback giver. This is the case for both verbal and non-verbal feedback, as well as multimodal feedback. Verbal feedback was provided more often under the condition when the speaker did not direct his gaze at the feedback giver.

Table 6.5: Proportion of positive feedback in relation to speaker's gaze in the analysed AMI meeting corpus (proportions in %).

	Total	Looked at by speaker	Not looked at by speaker	Unspecified
Verbal f.	16.6	11.0	3.1	2.5
Non-verbal f.	58.9	38.7	10.4	9.8
Combination	24.5	19.0	4.3	1.2
Total		68.7	17.8	13.5

6.2.2 Articulate feedback

Articulate feedback is expressed explicitly by means of utterances with a nontrivial syntactic form and semantic content. Clark calls such a type of feedback *assertion of understanding* (Clark, 1996). Articulate feedback acts often use general-purpose communicative functions. For example:

(59) 1. What do you mean by that?
2. Does this make sense?
3. Could you repeat this
4. I see what you mean

These dialogue acts in (59) are all feedback acts: (1) is a Set Question signalling that the speaker's has an interpretation problem; (2) is a Propositional Question where the speaker elicits feedback at the level of interpretation; (3) is a Request to repeat the previous utterance, because speaker's perception failed; and (4) is an Inform about the speaker's understanding.

This is, however, not the only means that a speaker has for an articulate reference to his (or his partner's) processing state. To signal successful perception of the previous utterance, speakers often repeat part of it. This phenomenon is sometimes called *implicit verification*. It is frequently used in spoken dialogue systems to allow the user to verify the correctness of the system's speech recognition, and gives the user the possibility to correct speech recognition mistakes on the fly. There is, however, nothing implicit about 'implicit verification'; the more

evidence of this kind is provided, the more explicit the speaker is about his processing state. For example:[6]

(60) F1: I'm above the chimney just now
G1: You are above the chimney and if you start going to in an angle about forty-five degrees

The speaker of G1 repeats most of the utterance F1 to signal his successful perception. This form of feedback can also be used for the purpose of winning some time. In this example the instruction-giver needs some time for orientation on the map and formulation of the next instruction (the repetition is almost twice as long as the original segment: 1201ms and 638 ms respectively).

In spoken human-computer dialogue repetitions can be very useful, but if they occur all the time the dialogue feels unnatural and too long. People mostly use more subtle ways to create confidence that the communication is successful. Repetitions can, for instance, be replaced by paraphrases. The speaker in this case signals not only successful perception, but also understanding. For example:[7]

(61) D1: There are zones frequencies as well as characters, different keypad styles
B1: Oh yeah regions and stuff

Another way to signal successful processing and move the dialogue forward is by using discourse markers. For instance, the discourse marker 'then' is used as such in the following example:[8]

(62) B1: Anybody anything to add?
A1: No
B1: That's the end of the meeting then

The discourse marker 'and' often signals that what is discussed up to now is processed successfully and that the upcoming speech will add new information or mark a transition to another discussion topic. Other discourse markers, such as 'well' and 'but', rather signal processing problems, mostly at the level of evaluation. For example:[9]

(63) A1: I think the option of the kinetic thing which means as long as you shake it like a watch
D1: But are people gonna wanna shake their movie controller

The multidimensional semantics of the most frequently occurring discourse markers is discussed in more detail in Section 6.4.

All expressions of articulate feedback may be accompanied by non-verbal signals. The most frequently occurring signal of positive feedback is the head nod, and of negative feedback the head shake and certain distinctive facial expressions. The next subsection discusses experiments carried out in order to understand how people interpret various types of head movements in dialogue.

[6] From the MapTask corpus - q1nc1.
[7] From the AMI meeting corpus - ES2002a.
[8] From the AMI meeting corpus - ES2002a.
[9] From the AMI meeting corpus - ES2002b and ES2002c.

6.2.3 Grounding by nodding

Feedback acts are closely related to *grounding*. To be successful, participants in dialogue have to coordinate their activities on many levels. In the speaker role, a participant not only produces utterances but also evaluates whether the addressee(-s) attend to, perceive, understand, and react to the speaker's intentions. An addressee's task is to attempt to understand the speaker's utterances, react to their intentions, and report on his processing. The coordination of the beliefs and assumptions of the participants is a central issue in any communication, the basic coordination problem being that of building shared or mutual beliefs out of individual ones. A set of propositions that the dialogue participants mutually believe is called their *common ground* (Clark and Schaefer, 1989), and the process of establishing and updating the common ground is called *grounding*. While common ground is not directly observable, grounding mechanisms are accessible through observable dialogue behaviour, e.g. evidence of understanding what is said in dialogue is provided by feedback acts. The nature of such evidence depends on the communicative situation. In face-to-face conversation, for example, participants may present evidence of grounding through body movements and gaze re-direction, while in telephone conversations only verbal and vocal signals are available.

Nonverbal means play an important role in the grounding process in face-to-face dialogue. For example, eye gaze is the most basic form of showing attention to what the speaker is saying, and head nods have a communicative function of acknowledgment signalling that the previous utterance was understood, without necessarily signalling acceptance (Clark, 1996). Goodwin (1981) notices that dialogue participants utilize both their bodies and a variety of vocal phenomena to show each other attention. For example, the speaker makes pauses and restarts his utterance when his gaze reaches a non-gazing recipient, or when late-arriving gaze of a recipient reaches a gazing speaker, or when recipient movements are noticeably delayed. Novick et al. (1996) found that the proportion of mutual gaze during conversational difficulties is greater at turn boundaries than within the turn. Nakano et al. (2003) observed that maintaining gaze on the speaker is interpreted as evidence of non-understanding, requesting additional information; by contrast, continued gaze on task-related objects (e.g. looking on a map) is interpreted as evidence of understanding.

All these findings suggest that nonverbal communicative means contribute especially to lower levels of grounding, signalling attention, perception and understanding of each other's communicative actions. As grounding may occur at any level of processing, one would expect evidence of grounding to also be provided at higher levels, such as evaluation and the adoption of beliefs. We show that this certainly happens in the case of complex nonverbal signs such as combinations of head nods, gaze re-direction and facial expressions. Such nonverbal evidence of higher-level grounding is observed in empirical data, and was found to be successfully recognized by multiple judges.

Information is transferred from one dialogue participant to another through belief creation (understanding) and belief transfer (adoption) (Bunt et al., 2007a). An utterance is understood by the addressee if he comes to believe that the preconditions of an intended dialogue act hold. For example, if A asks B a question then the understanding of A's question will be that B believes that A wants B to know some information, and that A assumes that B has this information available. The grounding of this question not only requires its understanding, but also evidence of believing. If B provides an answer to the question, then A may be expected to believe that B believes that the information he provides is correct, and B wants A to believe that the information provided by B is correct.

Speaker		Utterance								
B1	speech	but I th		I think	regardless we're we're aiming for	the under sixty five	or something			
	gaze	personD	personA		personD	personA	personC	personA		
	head									
	face									
	posture	working position								
A1	speech							Under sixty five	okay	That's a good start
	gaze	personB					table			
	head		short single nod		multiple short nods(5)			multiple long nods(4)		
	face									
	posture	working position					bowing			
D1	speech									
	gaze	personA		personB				table		
	head				multiple short nods(5)					
	face									
	posture	working position								
C1	speech							yep		
	gaze	personD	personA	personB					personA	
	head							long nods(2)		
	face							blinking		
	posture	working position								

Figure 6.1: Example of multimodal utterances from the AMI corpus.

To be sure that information is indeed transferred, a speaker needs evidence of correct understanding of his communicative behaviour and of being believed. In face-to-face interaction speakers receive such evidence through verbal and nonverbal expressions. The example in Figure 6.1 shows that different nonverbal and verbal expressions and their combination may convey different meanings. In this example, B says "but I th I think regardless we're we're aiming for the under sixty five". To come believe that *p* ('we are aiming for the under sixty five'), B should get evidence that A, C and D understand his utterance and believe its content *p*. The first head movement of speaker A in combination with gaze directed to B signals his understanding of speaker B's intention to have the turn; A's and D's multiple short head nods signal their understanding of B's intention to continue as a speaker. A's utterances 'Under sixty five', 'Okay' and 'That's a good start' accompanied by multiple short nods provide evidence of understanding (and positive evaluation) but not of adoption, since A offers that proposition for further debate. The evidence of understanding and adoption is provided by speaker C when he directs his gaze to B, performs long double nods (where the first one most probably indicates understanding) accompanied with single eye blinking and verbal 'Yep' to express agreement with B's inform. Thus, B believes that C believes that B believes that *p* and B weakly believes that C believes that *p*. In the grounding model of Bunt et al. (2007a), these beliefs may be strengthened by continuing dialogue when both have evidence that both know that both believe that p.

Therefore, as we see in the example in Figure 6.1, some evidence given nonverbally is about understanding, and its interpretation does not lead to belief transfer, whereas other nonverbal signals may be interpreted as successful belief adoption. In the next section we examine which types of nonverbal expressions and their combinations can be interpreted as adoption signals and which merely signal understanding. This will be investigated by means of perception experiments with multiple judges.

Types of nodding: Perceptual study

From the annotated AMI meeting data we randomly selected 60 video clips with 6 different speakers (3 male, 3 female). All six meeting participants were English native speakers. The duration of each clip was about 10 seconds and contained the full turns of the previous speaker and the current speaker. 16 naive subjects (4 male and 12 female, all between the ages of 20 and 40) participated in the experiments. They were given the task to answer the question whether they think that a participant *understands* the previous speaker or that he/she *agrees* with the previous speaker. Subjects had 10 seconds to react to each stimulus and were allowed to watch every video as many times as they liked.

Inter-subject agreement was examined using Cohen's kappa measure (Cohen, 1960). The judges reached a substantial overall agreement rating the stimuli (overall kappa 0.68). They recognized the signals of belief adoption better than those of understanding (kappa scores of 0.9 and 0.54 respectively). Next we determined nonverbal features that might be helpful for explaining why a participant's behaviour was interpreted as an expression either of correct understanding or of belief adoption. The following features were investigated:

- wording of an utterance;
- gaze (to person, table, slides, or averted);
- head movement, any or none;
- nods or jerks (any or none) and for these:

  - number of repetitions;
  - duration;
  - floor transfer offset;
  - speed (number of movements per second);
  - size (extra small, small, medium, large, extra large);
- eyebrow movement, any or none;
- eye shape change (e.g. blinking, widen, narrow), any or none;
- lips movement, any or none;
- hand movement, any or none;
- posture shift, any or none;
- some combinations of these features.

We performed Pearson's correlation tests and measured for each class label the correlations between the proportion of judges that chose this label and the features above. Table 6.6 presents the correlation results for the 'adoption' label (the correlation coefficient values for the 'correct understanding' label are the opposite ones). It is observed that if the dialogue participant combined head nods with verbal elements, especially the use of 'yeah', this was perceived by evaluators as a signal of belief adoption. Combination of 'uh-uhu' and head nods is more ambiguous; no significant correlation was observed. Signs of understanding are usually produced more silently. The speaker usually signals that he has understood the contribution without showing his acceptance or agreement. Understanding feedback utterances notably overlap the main speaker's utterance (average fto = -850ms); they are used frequently around utterance boundaries. Expressions of belief adoption, by contrast, are used around *turn* boundaries and may slightly overlap the main speaker utterance (average fto = -54ms). Head nods were mostly interpreted as adoption/agreement signals, and jerks (single backward head movement) as signals of understanding. The number of head nods positively correlates with the agreement interpretation: the more nods, the more probable that the speaker is adopting the partner's beliefs. Slow multiple head nods were also interpreted by most of the judges as signals that beliefs are adopted.

As for gaze pattern, when agreeing with their partners speakers exhibit certain regularities in the gaze behaviour that accompanies their head nods. They first look at the partner and avert their gaze near the end of the agreement phrase. Distinctive for agreement utterances were head nods in combination with lips movements, the speaker either flattening the lips (the mouth appears to be longer than usual in the horizontal plane, with lips compressed against the teeth) or smiling (lips corner-up and elongated). The test results also show that dialogue participants when expressing agreement with their partners often perform head nods together with eye blinking. Thus, head movements, which are diverse in form, speed, number of repetition, timing and accompanying verbal and nonverbal signs, convey different meanings and therefore play a different role in grounding processes.

In this study we used the DIT model of grounding in dialogue, which views information exchange as occurring through understanding and believing each other. We assumed that dialogue participants would provide different types of evidence to their partners if they merely *understand* the partner's intentions then if they also *adopt* the information provided. We studied several types of head movements that correlate with understanding and adoption, and investigated the features of understanding or adopting behaviours which are used to interpret

Table 6.6: Correlations between features and the proportion of votes for 'adoption'. (* differs significantly from zero according to two-sided t-test, $t < .05$)

Feature	Pearson's R
head nod(-s) + wording	.55* (p=0.000)
head nod(-s) + 'yeah'	.43* (p=0.000)
head nod(-s) + 'uh-uhu'	.2 (p=0.123)
duration	.17 (p=0.186)
floor transfer offset	.34* (p=0.07)
speed of movements	.22 (p=0.07)
size of movements	.027 (p=0.834)
number of repetitions	.25* (p=0.045)
head nod	.29* (p=0.02)
head jerk	-.29* (p=0.02)
gaze pattern 'person-averted'	.47* (p=0.06)
head nod(-s) + blinking	.25* (p=0.49)
head nod(-s) + eyebrows movement	.012 (p=0.925)
head nod(-s) + lips movements	.42* (p=0.001)
head nod(-s) + hand movements	.039 (p=0.762)
head nod(-s) + posture shift	-.16 (p=0.210)
fast single nod	-.13 (p=0.305)
fast multiple nods	.13 (p=0.32)
slow single nod	-.025 (p=0.847)
slow multiple nods	.37* (p=0.003)

these signals. We showed that dialogue participants use multiple signals and modalities to provide grounding evidence at different levels, and that conversational partners perceive and understand each other's intention more accurately when they can rely on multiple information sources. Also the interaction between different parameters of movements, such as number of repetition, speed and size of movements plays an important role when interpreting partner's behaviour. The importance of interaction effects between different parameters has been also emphasised when evaluating behaviour of animated agents. For example, in their evaluation study Hartmann et al. (2005) concluded that not only the technical implementation of individual parameters is important in order to achieve higher quality animation and better visibility of changes to the parameters, but also the interdependence of expressivity parameters, such as quantity of movement during a conversational turn, amplitude of movements, duration of movements, smoothness and continuity, dynamic properties of the movement and tendency to rhythmic repeats of specific movements needs to be reflected.

6.3 Turn organization

In the widely quoted study of Sacks et al. (1974) a model for the organisation of turn-taking in informal conversations has been proposed. The authors observed that conversations most often proceed fluently, that mostly one conversational partner talks at a time, that occurrences of more than one speaker at a time are brief, and that transitions from one turn to the next without a gap or overlap are very common. They reasoned that there must be an underlying system of turn-taking involved in conversations. They posited that during a conversation there

are natural moments to end a turn and initiate a new one, called Transition Relevance Places (TRPs), and formulated the following rules:

- If the current speaker (S) selects the next speaker (N) in the current turn, S is expected to stop speaking, and N to speak next.
- If S's behaviour does not select the next speaker, then any other participant may self-select. Whoever speaks first gets the floor.
- If no speaker self-selects, S may continue.

The generality of these rules makes them explanatory and applicable in many situations, but prevents them from being specific about the characteristics of speaker-selection techniques. At least two questions remain: (1) Which perceived behavioural aspects are used by people to estimate the locations of TRPs, and (2) Which aspects of communicative behaviour serve as signals to determine who is a potential or intended speaker of the next turn.

It would seem a plausible assumption that people use breaks of silence as a cue to know when to start a new turn. However, it has been found that these breaks are very unlikely to serve as primary TRPs or speaker-selection cues, because pauses between turns are often even shorter than pauses within turns (Cassell et al., 1999). Also, the duration of breaks between turns is often shorter than the time people need to mentally formulate a new utterance. These findings indicate that listeners anticipate TRPs in such a way that they already start to formulate a new utterance before the end of a current turn, and that they can predict precisely when to start their turn. It was observed that many turn transitions happen without temporal delays because a potential next speaker knows when and how a turn ends. People are able to predict turn endings with high accuracy using semantic, syntactic, pragmatic, prosodic and visual features (Ford and Thompson, 1996; Grosjean and Hirt,1996; De Ruiter et al., 2006; Barkhuysen et al., 2008, among others).

When participants of a dialogue can see each other the organisation is easier than when they can only hear each other. This suggests that they do not only make use of auditory information in speech like prosodic, syntactic, semantic, and pragmatic characteristics of utterances, but also draw on aspects of visible behaviour in order to project TRPs and to signal who will be the next speaker (e.g. Padilha and Carletta, 2003; Mazeland, 2003). We investigated several of these nonverbal behaviours of conversational participan to determine their turn-organisational functions.

6.3.1 Who is next?

While end-of-turn prediction has been studied extensively (see Ford and Thompson, 1996; Grosjean and Hirt,1996; De Ruiter et al., 2006; Barkhuysen et al., 2008), little research has been done on the prediction of who is a potential next speaker, and on next speaker self-selection behaviour. This is important when we deal with more than two participants in dialogue. Dialogue participants may just start speaking if they want to say something, but they often signal their willingness or readiness to say something. In other words, they perform certain actions to take the turn over. Speakers may signal that they want to have the turn when it is available (*turn taking*); that they want and are ready to have the turn when it is given to them (*turn accepting*); and that they want to have the turn despite the fact that it is not available (*turn grabbing*).

We studied the properties of a speaker's utterances that correlate with his turn-obtaining efforts in multi-party dialogue. Correlations indicate that two variables are related, but do not

measure cause; it cannot be concluded that signs which are correlated with turn-obtaining efforts are interpreted as such by communicative partners. To investigate this, we studied whether speaker changes really occur shortly after certain signals have been sent. We should also take into account, however, that a participant's wish to have the turn may be overlooked or ignored, and that he does not always get the opportunity to speak. Therefore, to obtain more certainty about utterance properties related to turn taking, we performed perception experiments where subjects judged the participant's turn-taking efforts.

Observation study

In the selected AMI data (2.400 functional segments), 412 segments were identified having a turn-initial function (17.2%) and 370 segments as having one of the turn final functions (15.4%).

We examined agreement between annotators in identifying and labelling turn management segments using Cohen's kappa measure (Cohen, 1960). Two annotators who were experienced in annotating dialogue and were thoroughly familiar with the tag set reached substantial agreement (kappa = .86) in identifying turn segments and assigning turn-management functions.

Results of the observation study

It was observed from the annotated data that meeting participants often indicate explicitly when they wish to occupy a sender role. More than half of all speaker turns were preceded by attempts to gain the turn, either verbally or nonverbally, or by combination of those (59%). 17.2% of all functional segments were found to have one of the turn-initial functions: 12% are turn-taking segments, 4.4% have a turn-grabbing function and 0.8% are turn accepts. Consider the following examples:[10]

(64) B: What did **you guys** receive? *(Turn Release)*
A1: $_{0.54}$ **Um**$_{(0.65)}$ *(Turn Take)*
A2: I just got the project announcement

(65) B1: Yeah brightness and contrast
D1: $_{-0.35}$ **Well**$_{(0.19)}$ *(Turn Grab)*
D2: $_{0.11}$ What we're doing is we're characterizing

(66) B1: That's something we'd want to include
B2: Do **you**$_{\text{(participant D is gazed)}}$ think? *(Turn Assign)*
D1: $_{1.82}$ **Uh**$_{(1.39)}$ *(Turn Accept)*
D2: Sure

The reasons to take the turn may be various. First, a participant may have reasons to believe that he was selected for the next turn by the previous speaker. This puts a certain pressure on him to accept the turn[11]. Second, a dialogue participant may want to make a contribution to the dialogue and believe that the turn is available. Finally, a dialogue participant may wish to have the turn while believing that it is not available, for example because he has a desire to express his opinion urgently, or he failed to process the previous utterance of another participant and needs immediate clarification, or he expects the current speaker to finish his utterance, and wants to make sure that he will be the next speaker.

[10] From the AMI meeting corpus - ES2002a.

[11] We did not observe turn refusing acts in our data. Turn refusal acts can be only performed non-verbally, since verbally in order to refuse the turn one would need to take it first.

Verbally, turn-taking intentions were mainly expressed by the following tokens: *um* and its combinations such as *um okay*, *um alright*, *um well* and *um yeah* (11.5% of all turn-initial segments); *so* (5%); *and* and combinations like *and so*, *well and*, also by *um and*, *uh and*, *and um*, *and uh* (7.9%); *well* (5.8%); *right* and combinations like *right so* and *right well* (7%); *uh* (5.6%); *okay* and *mm-hmm*/*uh-uhu* (5%); *alright* (2.8%); *yeah* or its repetition (15.7%); *but* (2%); *just* (1.2%); and repetitive expressions (e.g. *I.. I.. I.. would like*) (1.5%).

The majority of these tokens may serve several communicative functions is dialogue. For example, *'um'* and *'uh'* are known to be used as fillers both to stall for time and to keep the turn. Moreover, these tokens also occur in segments which are not related to turn management. For example, *'okay'* can be used as positive feedback or to express agreement. They also can be multifunctional expressing, for example, positive feedback and turn taking simultaneously. Previous studies, e.g. Hockey (1993) and Gravano et al. (2007), confirmed that the use of these cue phrases can be disambiguated by their position in the intonation phrase and their pitch contour.

We observed significant mean differences between turn-initial and non-turn-initial use of these tokens in terms of duration (turn-initial tokens being more than 115 ms longer); mean pitch (turn takings have $>$ 12Hz); standard deviation in pitch ($>$ 5Hz); and voicing (5% more voiced). As for the temporal properties of verbal turn-initial functional segments, the floor transfer offset (FTO) is between -699 and 1030 ms, where negative value means overlap and positive a gap between successive turns. Turn-grabbing acts have an FTO from -699 to -166ms; turn-accepting acts may also slightly overlap the previous segment and have FTO from -80ms to 136ms; turn-taking acts have the longest positive FTO (between 582 to 1030ms).

To assess the importance of nonverbal signs for identifying turn-initial segments, we conducted a series of correlation tests using the phi-coefficient. The phi measure is used to test the relatedness of categorical variables, and is similar to the correlation coefficient in its interpretation. Table 6.7 shows the correlation between segments annotated as having a turn-initial function and accompanying nonverbal signals.

Table 6.7: Nonverbal signals correlated to turn-initial segments (* significant according to two-sided t-test, $<$ .05)

(Non-)verbal signal	ϕ
wording (presence of tokens listed above)	.47*
any gaze redirection	.79*
gaze: direct-averted	.42*
gaze: direct($>$1 person)-averted	.61*
head movement	.05
hand/arm movement	.01
eye shape change + eyebrow movement	.15
any lips movement	.59*
half-open mouth	.39*
random lips movements	.28*
posture shift	.87*
working position-leaning backward/forward	.29*

Strong positive correlations were observed for gaze aversion, lip movements and posture shifts. Especially in multi-party conversations gaze plays a significant role in managing fluent turn transitions, more than in two-person dialogues, because of the increased uncertainty about

who will be the next speaker. In 11.8% of turn-initial segments the participant who wants to have the next turn gazes at more than one of the partners, most probably verifying their intention concerning the next turn. A dialogue participant who aims for the next turn first gazes at one or more partners, and averts his gaze shortly before starting to speak (44.1%). Comparable patterns were observed in previous studies (see e.g. Goodwin, 1981; Novick et al., 1996; Kendon, 2004).

Head movements are used for turn management purposes. In our data the intention to have the next turn was successfully signalled by repetitive short head movements (34.3% of turn-initial segments). In 11.8% turn-initial efforts were signalled by waggles (head movement back and forth and left to right). In 3.9% of turn-initial segments headshakes are signals of disagreement. Interestingly, however, head movements do not correlate significantly with turn-initial acts. By contrast, a combination of spoken signals like 'okay' or repetition of 'yeah' and multiple head nods are good signals of a participant's turn-obtaining intention (ϕ=.41, $p = .003$).

Hand and arm gestures that may be related to the participant's intention to have the turn were not observed frequently. We identified some shoulder shrugs that signalled uncertainty (3.5% of turn-initial segments) accompanied by head waggles and hand movements when a participant listening to the speaking partner suddenly moves his hand/fist away from the mouth (2%) or makes an abrupt hand gesture for acquiring attention (3.9%).

Generally, dialogue participants recognize an intention to take the turn successfully. In 60.8% of all the cases turn-obtaining efforts were acknowledged and the partner's wish to have the turn was satisfied. Participants who used more than one turn-initial signal or two modalities (e.g. combining head movements and posture shifts, or verbal and nonverbal signs) were more successful in obtaining the next turn. As for the remaining 39.2% it is difficult to judge whether the turn-taking efforts were interpreted as such by partners and ignored, or whether the signals were overlooked. Looking closer at gaze behaviour of meeting participants, our intuition is that in the majority of cases (65.2%) the turn-gaining efforts were most probably overlooked, because the participant was not gazed at by other partners. In another 34.8% of the cases, the participant's turn-gaining efforts were most likely ignored, since the partners did have direct eye contact. Nonetheless, since our analysis is based on the interpretation of annotators, this intuition could be wrong. To deal with this problem, perception experiments were performed which are reported next.

Perception study

Stimuli and procedure

Two series of perception experiments were performed to study whether naive subjects interpreted certain behaviour as signals to have the next turn. From the annotated data we randomly selected 167 video clips with 4 different speakers (2 male, 2 female). The following categories were considered:

1. a turn-initiating act is performed when the next turn became available;
2. a turn-initiating act is performed when the next turn was assigned to this participant;
3. a turn-initiating act is performed when the turn was not available but the participant needs: (a) to signal negative feedback on processing the partner's utterance; or (b) to elaborate the partner's utterance; or (c) to address the partner's suggestion; or (d) to clarify the partner's utterance; or (e) to shift the topic;

4. no turn-taking act is performed.

Two referees judged the selected clips for whether any turn initiating behaviour is observed. 52 stimuli, on which the judges fully agreed, were selected for further experiments: 4 of category 1; 4 of category 2; 36 of category 3; and 8 of category 4. The duration of each clip was about 10 seconds, containing the full turn of the previous speaker, and the recordings of the participant's movements and pause (if any) after the turn till the next turn starts. The subjects had 10 seconds to react to each stimulus. They were given the task to answer the question whether they think that a participant in question is performing any turn-initial act or not.

15 subjects (4 male and 11 female, all between the ages of 20 and 40) participated in one of the two sets of experiments: 9 subjects were asked to evaluate the video fragments without sound and 6 subjects evaluated the same fragments which were provided with sound. They were allowed to watch every video as many times as they liked.

Subject rating

Inter-subject agreement was measured using Cohen's kappa (Cohen, 1960). Table 6.8 shows kappa scores for each individual condition, for two class labels ('turn-initial act' and 'no turn-initial act' performed) and for two sets of ratings.

Table 6.8: Cohen's kappa scores for each class label for two sets of rating experiments

	without sound	**with sound**
turn take	.31	.65
turn accept	.20	.55
turn grab	.32	.43
no turn-initial act	.79	1.00
overall	.48	.64

Subjects reached moderate agreement judging whether a turn-initial act was performed if they could not hear what was said, relying only on the nonverbal information; they reached substantial agreement if they could hear what was said. Agreement is higher when a participant does *not* display any turn-taking efforts: (.79) when judging videos without sound, and 1.00 when sound was available. Among the turn-initial acts, turn grabbing has been identified with higher agreement than the others (.57, $t < .05$) under both conditions, most probably because participants produce distinctive facial expressions characterized by changing eye shape and eyebrow and lips movements, often accompanied by a head shake or waggle, additionally to other signals. The lowest agreement was found rating the turn-accept efforts of dialogue participants. This can be explained by the fact that participants to whom the next turn is assigned do not necessary perform any extra activity to indicate that they wish to be the next speaker. Raters who could hear what the other participants say reached much higher agreement than those who could not, so context information, such as the previous speaker's turn, is important for the perception of turn-taking behaviour, perhaps also because dialogue participants actually anticipate TRPs (de Ruiter et al., 2006), which makes it easier to perceive speaker-selection actions and to interpret turn-obtaining intentions.

Perceived properties of turn-initial acts

For explaining why subjects interpreted a participant's behaviour as having a turn-obtaining function we examined the following features: (1) gaze (directed, averted and a combination of those); (2) head movement, any or none; (3) hand gesture, any or none; (4) eyebrow movement, any or none; (5) eye shape change, any or none; (6) lips movement, any or none; (7) posture shift, any or none; and (8) some combinations of these features. Table 6.9 presents correlations for the conditions with and without sound.

We can conclude that nonverbal signals are important for recognizing speaker-selection intentions. A gaze pattern such as 'gazing at more than one person and then averting the gaze', and various types of lips movements and (half-)open mouth in particular, correlate positively with a turn-initial act and have strong negative correlation with non-turn-initial act).

Table 6.9: Correlations between features and the proportion of votes for each class label (without/with sound ratings). * differs significantly from zero according to two-sided t-test, $t < .05$

	ϕ **(without sound)**	ϕ **(with sound)**
turn-initial act		
gaze 'averted'	-.54*	-.44*
gaze 'direct(more persons)-averted'	.54*	.52*
head movement	.49	.25
head nods	.40	.28
hand gesture	.49	.21
eye shape change + eyebrow movements	.54*	.46*
(half-) mouth	.58*	.35*
lips movement	.44	.34*
posture shift	.41	.30*
'posture shift + head movement'	.34	.35*
'lips + head movements'	.57*	.39*
'eye shape change + head movements'	.47	.27
'eyebrow + head movements'	.46	.25
'gesture + head movements'	.44	.15
gaze 'direct-averted' + posture shift	.37	.34*
gaze 'direct-averted' + head movement	.55*	.40*
gaze 'direct-averted' + lips movements	.60*	.59*

A combination of head movements and other signals was perceived by judges as a turn-initial signal, e.g. a head movement accompanied by posture shifts and certain gaze pattern such as 'mutual gaze - averted' (strong positive correlation with turn-initial acts: .55, $t < .05$). Dialogue participants who use multiple signals or modalities are more successful in gaining the turn. Conversational partners are also more likely to perceive and understand the partner's turn behaviour when relying on multiple information sources.

6.3.2 Keeping the turn

Linguistic cues

Dialogue partners utilize a rich repository of linguistic, paralinguistic or nonlinguistic means to signal that they want to keep the speaker role.

No less than 46.4% of all turn-keeping events occur immediately after segments with one of the turn-initial functions, when the speaker just claimed the turn. For example:[12]

(67) B1: 1.92 Okay (1.24)
B2: 2.16 **Um** (1.4)
B3: What are we doing next?
B4: 0.64 **Uh** (0.48)
B5: 2.56 Okay 0.52
B6: **Um** (0.58)
B7: We now need to discuss the project finance

In (67) the speaker (B) signals several times that he wants to stay in the speaker role, by taking the turn and keeping it in B1 (note: long duration of *okay*) and by using filled pauses in B2, B4 and B6.

In meetings the speaker may want to keep the turn when he needs time to consult his agenda, notes or slides as, for example, illustrated in example (68)[13], B3.

(68) B1: Were gonna have to wrap up pretty quickly in the next couple of minutes
B2: 0.28 Um (0.28)
B3: **I'll just check we have nothing else**
B4: 0.16 Okay
B5: Anything else anybody wants to add about what they don't like about remote controls

The speaker can directly monitor the utterance that he is currently producing or preparing to produce, and notice that he needs some time for the utterance production. He signals to the addressee that his intention is to keep the turn but that he experiences some difficulties which require a bit of time, often for finding the right word. To this end he can use filled pauses, but also editing expressions like *sort of* or *kind of*.

Between-units turn keeping acts account for 43% of all turn-keeping segments. For example:[14]

(69) A1: Finding them is really a pain you know
D1: Mm mm
A2: **I mean** its usually quite small or when you want it right it slipped behind the couch or its kicked under the table

Tokens that construct turn-keeping acts are the longest, approximately 86 ms longer than the tokens within other functional segments.

As for prosodic features, the standard deviation in pitch is the most important one, because if it is high it indicates the speaker's hesitation and uncertainty that Turn Keeping segments are usually characterized by ($>$ 2-5Hz). The mean pitch has a statistically significant effect as well ($>$ 6Hz higher than by non-turn events, and $<$ 24-34 other turn events). Turn Keepers are the most unvoiced functional segments ($<$ 8-16%).

Non-verbal cues

For each segment we calculated the proportion of direct gaze (gaze directed to any of dialogue partner) and averted gaze (gaze directed elsewhere, e.g. to an object, looking up or aside)

[12] From the AMI meeting corpus - ES2002a.
[13] From the AMI meeting corpus - ES2002a.
[14] From the AMI meeting corpus - ES2002a.

expressed in milliseconds. This proportion was normalized by the length of segments, since duration of segments of various types may differ significantly, e.g. one-token turn or time management acts and task acts such informs. For turn-keeping acts, the proportion of directed gaze is larger than those of averted gaze (61.5 and 38.5 respectively); see Table 6.10. Compared to other functional segments, turn-keeping units have a larger proportion of averted gaze and a lower proportion of direct eye-contact. This is statistically a significant difference ($\chi^2(1)$ = 16.45, p $<$ .001). There is also a strong positive correlation between averted gaze and turn-keeping acts: ϕ = .96, p $<$ .001). This suggests that speakers who want to keep the turn avert their gaze rather than look at the partners.

Table 6.10: Proportion of directed and averted gaze (in %) during turn-keeping and other functional segments

Gaze type	Turn-keeping	Other segments
Directed	61.5	79.1
Averted	38.5	20.9

Table 6.11: Relative co-occurrence frequency of visible body movement during turn-keeping and other functional segments

	Turn-keeping	Other segments
Posture shifts	2.0	4.6
Head nods	8.8	41.9
Waggles	69.3	7.7
No head movement	91.2	58.1
Hand/arm gestures	5.4	4.7
Forehead muscles constriction	31.8	6.1
Narrowing eyes	29.5	2.8
Lips movement (e.g. pout,compress, biting)	36.7	6.8

As for head movements (see Table 6.11), certain types of movements are associated with turn-keeping acts. Waggles, random up, circle and aside movements were observed. These movements and turn-keeping acts have a strong positive correlation: ϕ = .84, p = .006.

As for posture shifts, there is no strong with turn-keeping acts ($\chi^2(1)$ = 3.04, p = .08; ϕ = .07, p = .08). Similarly, no significant difference were observed between the distribution of gestures during turn-keeping acts and acts of other type ($\chi^2(1)$ = 5.24, p = .12).

When a speaker keeps the turn for the purpose of memory search or conversational planning, facial activity often displays this. For instance, it was observed that speakers very often constrict their forehead muscles (ϕ = .95, p = .001) and narrow their eyes (ϕ = .92, p = .001), indicating that they are thinking. Lips movement like pout, compress, protrude, bite, muck, etc. usually indicate cognitive processing. There is a strong positive correlation between lips movements and turn-keeping acts (ϕ = .91, p = .002).

6.3.3 Giving the turn away

A speaker who is ready with his contribution may just stop talking and thereby signal that the next turn is available. Often, however, the speaker selects one particular partner for the next

turn, i.e. assigns the speaker role to an addressee (**Turn Assign**). Sometimes the speaker does not put any pressure on a particular addressee to take the turn, but indicates that anybody may continue (**Turn Release**).

Turn Assign

A speaker has basically two possibilities to indicate verbally to the addressee that he is selected for the next turn: to use the proper name of the intended next speaker, or to use the personal pronoun *you* referring to one particular person, for example:[15]

(70) What(0.18) what (0.12) do **you** think **Craig**?

(71) What did **you** (gaze direction: participant A) get?

These signals may be accompanied by gaze direction; head movements; posture shifts (turning to the addressee); or a pointing gesture. If *you* is used, to determine the addressee may be a non-trivial task, because it could be used in a *referential* or in a *generic* sense. Efforts have spent on automatic disambiguation of these two cases for a slightly different purpose, namely addressee detection (see Gupta et al., 2007). The most important features which have been used to discriminate between different uses of *you* are sentential features, e.g. previous and next tokens; part of speech information; current and/or previous dialogue act tag(-s); and the question mark feature, when available from transcription.

The speaker may assign the turn to the selected partner using only non-verbal signals in a situation where a certain turn-taking routine has been established. For example:[16]

(72) B1: First of all just to kind of make sure that we all know each other
B2: I'm Laura and I'm the project manager
B3: Do you want to introduce yourself again
A1: Hi I'm David and I'm supposed to be an industrial designer
B4: Okay
B5: **turns to participant D; turns head to participant D + head nod; gaze direction: participant D**
D1: I'm Andrew and I'm our marketing expert

About 4.6% of all functional segments in the data have the communicative function Turn Assign. Turn-assigning acts are difficult to recognize automatically, since calling participants by name happens seldom and *you* is ambiguous, e.g. it may also refer to the whole group of participants which in this case would constitute a Turn Release act, and the information about non-verbal behaviour of the speaker is not always available to the system. Turn-assigning acts have some distinctive prosodic properties which can help their successful recognition, see Table 6.14. There are statistically significant differences in mean pitch: turn-assigning acts have a higher mean pitch ($>$ 26-30Hz) than non-turn events and then turn-keeping segments, and a lower mean pitch than turn-releasing acts ($<$ 11 Hz). Turn-assigning acts are more voiced ($>$ 9-16 voiced frames) than non-turn events and turn-keeping segments; no significant difference was found with turn-releasing acts. Turn-assigning acts are louder than immediately preceding non-turn event ($>$ 5dB). Finally, their speaking rate is lower than that of non-turn events ($<$ 1 syllable per second).

[15] From the AMI meeting corpus - ES2002a.
[16] From the AMI meeting corpus - ES2002a.

Releasing the turn

About 1.3% of all functional segments have the communicative function Turn Release. The most important features to successfully identify turn-releasing segments are a specific token and its surroundings (e.g. bi- or trigrams) and dialogue context, e.g. bigger segment they occur within, as well as tag of the previous functional segments. Turn-releasing segments in English contain the following tokens:

- *anybody*, *anything* or *any*, for example: '**Anybody** anything to add?', '**Anything** else to say at all?', '**Any** thoughts on that at all
- *everybody*, for example: 'Is that what **everybody** got?
- *we* or *all*, for example: 'Shall **we** make the decision?', '**All** ready to go?
- *you* referring to the group, for example B3 in 72

Prosodically, turn-releasing acts are characterised by a high mean pitch (as well as minimum and maximum pitch): 37-41 Hz higher than non-turn events and turn-keeping acts respectively, and 11Hz higher than turn-assigning. Effects of intensity was found as statistically significant ($>$ 3-4 dB). Moreover, turn-releasing acts are as a rule more voiced than non-turn segments and turn-keeping acts ($>$ 9-17%) and equally voiced as turn-assigning ones.

6.4 Discourse structure

Acts for opening and closing a dialogue and for topic management are responsible for structuring the interaction. Dialogue structuring acts are based on the speaker's view of the present linguistic context, on his plan for continuing the dialogue, and on an assumed lack of clarity of the topical structure of the linguistic context for the addressee(-s) (Bunt, 1996). DIT^{++} defines three communicative functions for discourse structuring: Topic Introduction, Topic Shift and Topic Shift Announcement. Topics may be introduced or changed in various ways.

Quek et al. (2000) noticed that gestures, postures and gazes frequently appear to mark the topic of the discourse. Cassell et al. (1999) proved empirically that the gaze behaviour of the speaker is directly related to discourse coherence and is a good indication of the new information in the utterance (Rheme - Theme relation). They noticed that the speaker usually has direct eye contact with interlocutors when producing new information, mostly a new discussion topic. Hand palm presentation gestures are typically used in association with passages in the verbal discourse which serve as an introduction to something the speaker is about to say (see Kendon, 2004).

Quek et al.(2000) noticed that a participant who shifts from one topic to another, shifts in posture, e.g. leaning forward and turning to the potential interlocutor. We observed that in 44.4% of all topic shifts the speaker changed the position of his upper-body, mostly leaning to the addressee. In 50% of all topic shifts the addressee shifts his posture as well, mostly positioning his upper-body towards the speaker.

Frequently occurring cues indicating discourse, topic and argumentation structure are *discourse markers*. Discourse markers have been studied for their role in the organization of discourse structure in larger texts (Mann and Thompson, 1988), in argumentative dialogues (Cohen, 1984), in interviews (Schiffrin, 1987; Hirschberg and Litman, 1993) and in dialogues that are highly interactive in nature and are characterized by rapid turn switching among participants, such as task-oriented dialogues (Heeman and Allen, 1999) or meeting conversations

(Popescu-Belis and Zufferey, 2006). In dialogue, discourse markers play an important role in establishing boundaries between dialogue units and in indicating the communicative functions of such units.

Discourse markers: multidimensional semantics

Studies of the functionality of discourse markers make use of various sets of underlying relations in discourse. Discourse particles such as 'well', 'so', 'but', 'and', etc., are traditionally acknowledged to have the status of a discourse marker. It has been argued, however, that signals related to grounding (e.g. 'okay', 'uh-uhu'), to timing in speech production (e.g. filled pauses), and to turn management (e.g. 'uh') should also be included in the class of discourse markers (e.g. Swerts and Ostendorf, 1997; Swerts, 1998; Louwerse and Mitchell, 2003).

One aspect of the meaning of discourse markers is that they may not only have a variety of semantic functions, but that they may also have several functions simultaneously – their *multi-functionality*. Schiffrin (1987) and Hovy (1995) argue that several parallel structures underlie coherent discourse, and argue that an adequate description of discourse requires at least four distinct structural analyses: semantic, interpersonal/goal-oriented, attentional/thematic, and rhetorical. They notice that discourse markers may simultaneously have roles in each of these structures, e.g. the discourse marker *and* may 'coordinate ideas' and 'continue a speaker's action'.

While these approaches seem to apply very well to the analysis of the meaning of discourse markers in dialogue, they have escaped comprehensive and formal description. The DIT framework makes it possible to compute a 'multidimensional' semantics of discourse markers by relating multiple context update operators to context models which include, besides information states of the usual kind (beliefs and goals related to a task domain), also a dialogue history, information about the agent's processing state, beliefs about dialogue partners' processing states, information and goals concerning the allocation of turns, and so on.

The interpretation of a multifunctional stretch of communicative behaviour corresponds to updating the context models of the communicating agents in multiple ways, combining the effects of each of the component functions. For example:[17]

(73) B1: Anything else what anybody wants to add about what they don't like about remote controls
A1:$_{0.48}$ **And** you keep losing them

Since it answers B's Set-Question B1, utterance A1, which includes the discourse marker *and*, updates the context model of participant B with the information that:[18]

(74) (1) *A believes that B wants to know what A does not like about remote controls*;
(2) *A believes that B believes that A knows what A does not like about remote controls*;
(3) *A believes that A knows what he does not like about remote controls*; and
(4) *A believes that B made the turn available.*

[17] From the AMI meeting corpus - ES2002a.
[18] For a formal representation of updates in participants' information state see Morante, 2007.

Table 6.12: Distribution and observed multifunctionality of discourse markers.

DM	Occurrence	Multifunctionality	GP	Auto-F.	Allo-F.	Turn M.	Time M.	DS	Contact M.	OCM
and	214	2.6	elaborate suggest exemplify explain	pos. evaluate pos. execute	pos. execute	take grab keep	stall	topic shift		
because	37	1.2	justify explain			keep	stall			
but	87	1.9	disagree correct explain warning	pos. evaluate neg. execute	neg. execute	take grab keep				
I mean	41	2	elaborate			keep				retract
like such as	70	1.7	exemplify			keep	stall			
oh	31	2	answer	pos/neg. execute		grab accept	pause	topic shift		error signal
so	226	2	conclude suggest elaborate	pos. execute		take grab keep give	stall	topic shift open pre- close	indi- cate	retract
then	45	1.9	instruct elaborate suggest	pos. execute		take grab keep	stall	topic shift		
well	63	2.1	disagree correct explain	pos. execute	neg. execute	grab take accept keep	stall	topic shift		retract
you know	84	2.3			check elicit	give keep	stall		check	retract

Thus, the simultaneous performance of the turn management and feedback acts through the use of A1, in particular of *and*, constitutes the multidimensional interpretation of what A says.

We found that discourse markers are used:

1. as indicators of rhetorical relations between dialogue acts;
2. as indicators of feedback dependency relations between functional segments;
3. as full-blown dialogue acts (without explicit semantic content), e.g. as a Turn Take act.

This means that discourse markers can have two kinds of meanings: as a dialogue act, i.e. as a context update operator, and as an element that contributes to the determination of the relations between dialogue acts and functional segments.

Table 6.12 lists the most frequent discourse markers (DMs) identified in our corpus with their absolute frequency, gives an overview of their observed multifunctionality and lists the observed communicative functions.

Note that all DMs may serve more than one communicative function. *And* is the most multifunctional discourse marker in our corpus, and *because* the least multifunctional one. *Because* mostly prefaces Informs with the rhetorical functions Justify or Explain, and is used in 2.4% of all occurrences to simultaneously perform Turn Keeping and Stalling acts. All discourse markers except *'you know'*, preface general-purpose functions (often in Task, Discourse Structuring, or Feedback dimensions) indicating various rhetorical relations, and may perform dialogue acts addressing some dimensions simultaneously. The latter pattern is observed for 50.7% of all studied DMs.

A discourse marker may also perform full-fledged dialogue acts addressing more than one dimension simultaneously. This is often the case for Turn Management in combination with Feedback, Time Management, Discourse Structuring or Own Communication Management (27.7% of all discourse markers are observed to be used in this way). It was noticed that at most 3 dialogue acts are performed by one discourse marker in a given context, e.g. as feedback, turn and time management acts.

A third pattern of DM use, 18.2% is as a single dialogue act, e.g. a turn taking act or a feedback act. In the rest (3.4%) discourse markers are part of a general purpose functions and do not perform a dialogue act on their own.

Different uses of discourse markers require successful recognition based on observable features like differences in prosody and features from surrounding lexical material, such as frequent word collocations (see results on automatic recognition of the discourse markers *'like'* and *'well'* reported by Popescu-Belis and Zufferey (2006), and for *'and'* reported by Petukhova and Bunt (2009b)). We observed significant mean differences for both raw and speaker-normalized features in terms of duration (DMs are almost twice as long as non-DMs: 327ms and 217ms respectively, and Stallings and Keepings acts are even longer: 585ms); initial pause (no or a negligible pause before non-DMs, and initial pauses before DMs between 59 and 228ms); mean pitch (*and* as DM has higher mean pitch: $>$ 12Hz). Preceding and following tokens as features also have high information gain.

And is the most frequently used 'classical' discourse marker in our corpus. In 45.5% of its occurrences, *and* is used as a discourse connective and in the rest as a propositional connective. Differentiating between *and* as non-DM and DM is important for segmentation purposes. Used in clause-initial position or as an autonomous segment, *and* as DM so to speak brackets segments and helps define their boundaries.

And as a discourse marker may have various and multiple communicative functions in dialogue. According to Schiffrin (1987), *and* as discourse marker 'coordinates idea units' and 'continues a speaker's action', and it has these roles simultaneously. In terms of DIT, *and* signals rhetorical relation such as Elaboration between dialogue acts in different dimensions, often in the Discourse Structuring dimension to preface Topic Shift), and it has some dimension-specific (DS) functions, often in the Turn Management dimension: Turn Take, Grab, or Keep. For examples and further details see Petukhova and Bunt (2009b).

6.5 The role of nonverbal behaviour

In previous sections we have already seen that multimodal behaviour is important for expressing and recognizing speaker's intentions. In this section we focus on what difference it makes when we consider linguistic behaviour only, and when we take non-verbal behaviour into account. To establish the role that nonverbal behaviour plays in dialogue, we conducted two annotation studies where annotators were asked to annotate dialogues with the DIT^{++} tagset: (1) using only speech transcription and sound; (2) using speech transcription, sound and video provided with transcriptions of nonverbal signals (gaze, head, facial expression, posture orientation and hand movements).

We examined agreement between annotators in labelling communicative functions using Cohen's kappa measure (Cohen, 1960). Two experienced annotators reached substantial agreement (kappa = .71).

We compared the annotations with respect to the number and nature of (1) functional segments identified; (2) communicative functions altered; and (3) communicative functions assigned to single functional segments.

In both studies we used two scenario-based dialogues with a total duration of 51 minutes from the AMI corpus, transcribed as discussed in Section 6.1.

The analysis showed that nonverbal communicative behaviour may serve four purposes:

1. emphasizing or articulating the semantic content of verbally expressed dialogue acts;
2. supporting the communicative functions of synchronous verbal behaviour;
3. performing separate dialogue acts in parallel to what is contributed by another participant;
4. expressing a separate communicative function in parallel to what the same speaker is expressing verbally.

Full-fledged dialogue acts

When the visual modality is taken into account in addition to the speech, about 20% more functional segments are identified: 1,917 versus 2,396. The 479 new functional segments, which have only nonverbal components, form a single full-blown dialogue act or multiple dialogue acts. These acts mainly address auto-feedback (68.5% : 3.2% negative, 65.3% positive). Signs of feedback notably overlap the main speaker's utterance (average 850 ms).

The nonverbal expressions used, include gaze direction, head movements (nods, jerks, waggles and shakes), posture shifts (leaning forward, backward or aside, shifting one's weight in a chair), especially in relation to attention, and facial expressions (e.g. lowering or rising eyebrows, lips movements, smile, blinking).

Of the dialogue acts performed nonverbally 24.7% are used for the purpose of managing the allocation of turns. Hand and arm gestures that may express the intention to have the turn include hand movements when a participant listening to the speaking partner suddenly moves his hand away from the mouth or makes an abrupt hand gesture. Various types of upper-body posture shifts are also often used as turn-initial signals; leaning forward, backward or aside, producing random shifts (shifting one's weight in a chair).

4.8% of all nonverbal acts has the communicative function of *stalling* (time management). Gaze aversion within an utterance, head waggles, and various types of self-touching (scratching, or holding the back of the neck or head with the open palm and rubbing the cheek or side of the neck) were interpreted as stalling signals.

About 2% of the nonverbally performed dialogue acts are used for dialogue structuring. Topic shifts are announced by raising a hand or a finger and palm-down gesture. Establishing mutual gaze and positioning the upper body in the working position, or breaking mutual gaze and leaning backward, respectively, are used for opening and closing the dialogue.

Communicative function alteration and specification

In a number of cases the communicative functions assigned to speech segments were corrected after annotators got access to visual signals. Mostly, this concerned an adjustment of the level of feedback, e.g. from understanding to evaluation or execution (6.8%). Petukhova and Bunt (2009a) noticed that participants in dialogue provide different types of evidence to their partners if they merely understand the partner's intentions than if they also adopt the information provided (positive execution feedback). It was shown that dialogue participants use multiple signals and modalities to provide evidence of grounding at different levels, and that conversational partners perceive and understand this more accurately when they can rely on multiple information sources. While simple head nods were perceived as a signal of successful understanding, more complex expressions, such as a combination of multiple slow head nods with lip movements and blinking, were perceived as signals of belief transfer (adoption).

Kendon (2004) observed that nonverbal acts which are not part of the propositional or referential meaning of the utterance may have modal functions, e.g. indicating whether the speaker regards what he is saying as a hypothesis or as an assertion. About 47% of all functional segments in our data are modalized (34.5% uncertain, 12.6% certain). A degree of certainty can be expressed verbally as well as nonverbally. Table 4.3 in Section 4.4 gives an overview of observed expressions.

Nonverbal expressions may reveal the speaker's attitude towards the addressee(-s), towards the content of what he is saying, or towards the actions he is considering to perform, and his emotional state. Pavelin (2002) calls these nonverbal expressions *modalizers* or *modal* gestures. We observed the following attitudes and emotions in our data: *thinking* or *reflecting*, *surprised*, *questioning*, *confused*, *amused*, *sceptical*, *interested*, *disappointed*, and *guilty*. Attitudes and emotions are mostly communicated by face.

Multifunctionality in multimodal utterances

A verbal functional segment has on average 1.3 independent communicative functions (also confirmed in Bunt, 2009b), whereas we observed that using information from all modalities gives 1.4 independent functions per segment on average. Table 6.13 presents the relative frequency of co-occurrences of multiple functions in various dimensions.

Table 6.13: Co-occurrences of communicative functions across dimensions in % for verbal expressions only and when including nonverbal expressions. (Read as follows: percentage of segments having a communicative functions in the dimension corresponding to the column, which also has a function in the dimension corresponding to the row.) Only functions are considered which are explicitly expressed.

	type	Task	Auto-F.	Allo-F.	Turn M.	Time M.	DS	Contact M.	OCM	PCM	SOM
Task	verbal		1.1	0.1	5.6	2.6	0.3	0	4.3	0.3	1.5
	multimodal		1.2	2.7	8.5	3.4	0.3	0	4.6	0.3	1.5
Auto-F.	verbal	0.5		0	12.7	0.5	0.3	0	0	0	0
	multimodal	0.7		0	15.5	2.6	3.1	0	0	0	0.5
Allo-F.	verbal	0	0		23.7	1.2	0	0	0	0	0
	multimodal	3.3	0		23.7	1.5	0	0	15.4	5.1	0
Turn M.	verbal	39.3	6.2	1.8		49.6	0.7	0	2.5	0	0.4
	multimodal	40.8	12.2	6.0		60.6	1.1	0.3	5.9	0.7	0.7
Time M.	verbal	34.6	0.5	0	9.1		0	0	0	0	0
	multimodal	41.7	3.5	11.2	9.7		0.5	0	4.2	1.4	0.6
DS	verbal	1.7	0	0	6.7	0		0	0	0	0
	multimodal	6.8	6.8	0	20.9	1.7		0	1.7	0	8.4
Contact M.	verbal	0	0		18.2	0	0		0	0	0
	multimodal	0	0	0	18.2	0	0		0	0	0
OCM	verbal	77.9	0	0	6.5	0	0	0		0	0
	multimodal	80.9	0	5.4	6.5	8.0	0.9	0		0	0
PCM	verbal	0	0	0	27.3	0	0	0	0		0
	multimodal	0	0	18.2	27.3	0	0	0	0		0
SOM	verbal	0.9	0	0	1.2	0	13.9	0	0	0	
	multimodal	0.9	1.2	0	8.3	1.2	13.9	0	0	0	

Although the average number of functions per segment does not differ much, multimodality is significant for enabling the multifunctionality of utterances in some dimensions. Nonverbal communicative acts are very often concerned with feedback and other interaction management dimensions. For example, speech-focused movements accompanying relatively unpredictable content words (e.g. iconic gestures during lexical search), and body-focused movements (e.g. searching for an elusive word in memory) normally indicate that the speaker needs some time to formulate an utterance, and is therefore stalling for time, but would like to keep the turn. Pauses near the beginning of an utterance can have the function of contact check, requesting attention. Speakers often make short pauses until the gaze of a recipient has been obtained and secured.

Nonverbal signals also add to utterances addressing the Task dimension some functionality in other dimensions; for example, as already noticed, gaze direction can have a turn management function, which may be additional to a task-related function.

Interesting nonverbal behaviour was observed with respect to speaker speech production and editing (own communication management). When the speaker's gaze reached a non-gazing participant or the partner's gaze arrived later with some delay, the speaker often restarted or retracted his utterance, indicating by this that he wishes to gain the addressee's attention. Such behaviour is multifunctional in the sense that the speaker signals that he corrects or retracts his utterance and by doing this intends to elicit feedback from his interlocutors.

Articulating semantic content

About 39% of all transcribed nonverbal signals neither contribute to the communicative function of a verbal utterance nor form a full-fledged dialogue act on their own. Nonverbal signals are often used for articulating the semantic content of a dialogue act.

First of all, nonverbal expressions are used to mark new, important information. As such they reinforce verbal communication and allow to accentuate or emphasise words or ideas. To stress the importance of information that the speaker is providing he can use beat gestures, which are known to accompany important information, as well as eyebrow movements to indicate where the focus of the addressee's attention should be positioned. Along with hand and eyebrow movements speakers often use head nods for emphasis coinciding with the most prominent words in an utterance. For example:[19]

(75) wording: I'm gonna do an opening
head: .....nod

Iconic, metaphoric and pantomimic gestures were observed which form part of the semantic content or specify the semantic content of an utterance, as illustrated in (76) - (78) respectively.[20]

(76) wording: The younger group of people would want smaller
hand: .....size(both hands)

(77) wording: Then we'll move into acquaintance including a tool training exercise
hand: .....semi-sphere..

[19] From the AMI meeting corpus - ES2008a.
[20] From the AMI meeting corpus - ES2008a.

(78) wording: Then we've moved to age group twenty five to thirty five
hand: .....away-motion.....

For detailed studies of the contribution of gestures to semantic content of an utterance see McNeill (1992) and Kendon (2004).

6.6 Summary

This chapter discussed the expression of other than task-related intentions, which are behind those communicative actions that are meant to manage the dialogue. We performed a variety of observational, statistical and perceptual tests in order to identify those properties of multimodal utterances that are potentially used by addressees to recognise the speaker's intentions. We have seen that feedback and turn management functions can be expressed verbally using much the same linguistic material. For example, *okay* can be used as a feedback check question, or to give positive feedback, or to express agreement; it can also be multifunctional, expressing, for example, positive feedback and turn taking simultaneously. Previous studies, e.g. (Hockey, 1993) and (Gravano et al., 2007), confirmed that the use of such expressions can be disambiguated in terms of position in the intonation phrase and pitch contour. We found that there are prosodic, temporal and durational differences between such multifunctional and polysemic tokens, see Table 6.14. We noted the importance of distinguishing between discourse marker use and non-discourse marker use, and found significant mean differences for both raw and speaker-normalized features in terms of duration, pause before the token in question, pitch features, voicing and speaking rate. Similarly, use as turn-taking acts can be distinguished from non-turn-taking use, as well as between turn management acts of different types.

One of the main conclusions with respect to the features of dialogue utterances is that addressees who use signals from multiple modalities are more successful in recognising the speaker's intentions, and that speakers who combine multiple modalities to convey their intentions can be more confident that their intentions will be recognised.

We investigated the role of nonverbal signals in dialogue more generally. Besides serving as full-fletched dialogue acts, nonverbal signals may emphasise or articulate the semantic content of dialogue acts; they may support the communicative functions of synchronous verbal behaviour; they may express a separate communicative function in parallel to what the same speaker is expressing verbally; and they may qualify a communicative function with respect to the speaker's certainty and sentiment. We noticed that, in a face-to-face setting, nonverbal signals accounted for 20% of the total number of functional segments and dialogue acts.

The observed linguistic and nonverbal expressions of dialogue acts of different types, summarised in Table 6.15 for feedback and turn management, can be used not only in the recognition of these acts, but also when generating natural dialogue behaviour, for example, by embodied agent systems.

Table 6.14: Prosodic features for frequently occurring tokens with different communicative functions; 95% confidence interval (Turkey post hoc test); *significance level $\alpha < .05$, one-way ANOVA test. (DM = discourse marker; fb = feedback; turn = Turn Management)

Token (functions)	Duration (in ms)	Initial pause (in ms)	Mean pitch (in Hz)	Standard deviation pitch (in Hz)	Fraction unvoiced /voiced frames (in%)	Intensity (in dB)	Speaking rate (in syl/sec)
and (DM/no)	289 367*	60 228*	142 156*	12 22*	22 28*	50 52	4.6 5.7 *
because(DM/no)	616 728*	-122 225	93 195	2 75	48 64*	41 50	5.4 6.5*
like(DM/no)	310 389*	-10 59	126 146	14 31*	40 49*	48 51	6 7.3*
so(DM/no)	297 355	12 656	152 172	22 35	47 52	51 53*	4.2 5.3*
well(DM/no)	281 350*	36 409*	152 176*	19 33*	30 38*	52 54*	5 6*
Um(Take/no)	202 - 264*	744 - 939*	4 - 17*	2 - 6	6 - 12*	50 - 54*	0.5 - 3*
Um(Accept/Take)	109 -154	365 -1026	8 - 66	8 -26	11 - 12	50 - 53	0.5 - 2
Um (Grab/Take)	93 201*	-1159 - -870*	6 -17	1 -12	2 - 7*	50 - 52	3.1 - 4.7*
Okay (fb/fb+turn)	296 - 1074	- 521 - 1192	43 - 89	22 - 42	1 - 46*	51 - 54	4.5 - 4.9
Okay (fb check/fb)	262 - 417	1119 - 1120	67 -126*	69 - 98*	21 - 44	51 - 52	2.4 - 9.2*
Okay (accept/fb)	273 - 439	1588 - 1595	91 -155	10 -130*	36 - 47	48 - 50	4.2 - 8.5
Yes (agreement/fb)	123 -293	- 660 - 1460*	56 - 115	10 - 12	19 - 48	53 - 62	4.6 - 5.0
Yes (answer/fb)	176 - 235	283 - 911*	9 -86*	4 - 36*	19 - 25	54 - 56	1.2 - 3.2
You (turn/no)	123 - 153*	0 - 257	112 -252*	11 - 22	0 - 18	24 - 74*	0.9 - 3.3*
You (assign/release)	190 - 220*	0 - 66	252-304*	38 - 42	67 - 75	60 -68*	9.1 - 10.5*

Table 6.15: Linguistic and non-linguistic expressions of feedback and turn management acts

Dialogue act	Verbal elements	Non-verbal elements				
		Gaze	Head	Hand	Face	Posture
Pos. Feedback						
attention	*'Uh-uhu'* *'Mm-mhm'*	speaker-directed	to speaker; slightly aside; short nod(-s)		forehead:constricted eyebrows:lowered eyes:narrowed mouth:half-opened	any shift
perception	repetition	speaker-directed	to speaker short nod(-s)			to speaker
interpretation	paraphrase yeah/yes	speaker-directed	short nod(-s) jerk		eyes:blinking lips: purse	forward aside
evaluation	discourse markers	directed-averted	long nod(-s)		eyes:blinking lips: corner-up, elongate or open	to speaker
Neg. Feedback	verification questions DM: *'well',' but'* repetitions	speaker-directed uncertainty:averted	shake; waggle	raising shoulder	forehead: constricted eyebrows: lowered, pulled together eyes:narrowed lips: compressed	to speaker or away
Turn Take	*'Um/Uh'* *'Okay/Right'* *'Well/And'*	direct-avert mutual-avert	short nod to partners		eyebrow:raise lips: half-open, random movements	any shift
Turn Keep	*'Um/Uh'* editing expressions DM: *'well'*	averted	waggles circle aside	adaptors iconic hold	forehead:constricted eyes: narrowed lips: pout	
Turn Give	proper name referential *'you'*	direct	deictic nod	deictic		to addressee

Chapter 7

Dialogue act recognition

This chapter is concerned with the automatic recognition of dialogue acts. We focus on the question of how the intended (multi-)functionality can be recognized based on observable behavioural features in a data-oriented way. We discuss and examine incremental token-based approach to dialogue utterance interpretation. A token-based approach combining the use of local classifiers, which exploit local utterance features, and global classifiers which use the outputs of local classifiers applied to previous and subsequent tokens, is shown to result in excellent dialogue act recognition scores for unsegmented spoken dialogue. This can be seen as a significant step forward towards the development of fully incremental, on-line methods for computing the meaning of utterances in spoken dialogue.

Introduction

The recognition of the intentions encoded in user utterances is one of the most important aspects of language understanding for a dialogue system. The state-of-art in dialogue act recognition is to use all available information sources from multiple modalities, including: (1) linguistic information, that can be derived from the surface form of an utterance: lexical and collocational information; (2) perceptual information from multiple channels available to dialogue participants, including acoustic and prosodic properties of utterances as well as information from visual and other modalities; (3) contextual information obtained from the preceding dialogue context and dialogue structure, as well as global context properties like dialogue setting, knowledge about dialogue participants, and domain knowledge.

Traditional models of language understanding for dialogue systems are pipelined and modular, and operate on complete utterances. Typically, such a system has an automatic speech recognition module, a language understanding module responsible for syntactic and semantic analysis, an interpretation manager, a dialogue manager, a natural language generation module, and a module for speech synthesis.

This chapter is based on Petukhova and Bunt (2011). The work reported in this conference paper was done by me in close cooperation with the co-author. This particular approach to dialogue act recognition has been inspired by the work of my colleagues Piroska Lendvai and Jeroen Geertzen.

In a pipeline architecture, the modules operate one by one in a sequential manner. The output of each module is the input for another. Thus, the interpretation manager that is responsible for dialogue act classification gets from the previous modules segmented dialogue utterances that have been syntactically and semantically parsed. This means that roughly speaking dialogue act recognition is performed in two steps: first, speech input is *segmented* and subsequently *classified* with dialogue act information.

Various machine learning techniques have been applied successfully to natural-language based dialogue analysis. For example, techniques based on n-gram language modelling were applied by Reithinger (1997) to the Verbmobil corpus, with a reported tagging accuracy of 74.7%. Hidden Markov Models (HMM) have been tried for dialogue act classification in the SWITCHBOARD corpus (Stolcke et al., 2000), achieving a tagging accuracy of 71% on word transcripts. Another approach that has been applied to dialogue act recognition, by Samuel et al. (1998), uses transformation-based learning. They achieved an average tagging accuracy of 75.12% for the Verbmobil corpus. Keizer (2003) used Bayesian Networks applying a slightly modified version of DAMSL with an accuracy of 88% for backward-looking functions and 73% for forward-looking functions in the SCHISMA corpus.[1] Lendvai et al. (2004) adopted a memory-based approach, based on the k-nearest-neighbour algorithm, and report a tagging accuracy of 73.8% for the OVIS data described in Section 4.5.1.

Apart from using different techniques, these approaches also differ with respect to feature selection strategies. Some approaches rely solely on the wording of an input utterance, using n-gram models or cue-phrase, e.g. Reithinger (1997) and Webb et al. (2005). Others successfully integrate prosodic features that facilitate accurate dialogue act recognition, e.g. Shriberg et al. (1998); Jurafsky et al. (1998a); Fernandez and Picard (2002); Stolcke et al. (2000). Again others combine the predictions derived from the utterance and its context, e.g. Keizer (2003); Stolcke et al. (2000); Samuel et al. (1998); Lendvai et al. (2004).

It is unlikely, however, that humans processing dialogue utterances wait every time until they get the complete utterance before trying to interpret what it is about and what type of dialogue act is performed. Evidence from the examination of transcripts of spoken conversations and from psycholinguistic experiments suggests that interpretation starts long before the complete utterance is constructed. Our observations of natural dialogue behaviour reported in Chapter 6 showed that humans process dialogue contributions *incrementally*, and are often able to anticipate the end of the utterance. Dialogue phenomena such as backchannelling (providing feedback while someone else is speaking), the completion of a partner utterance, and requests for clarification that overlap the utterance of the main speaker, illustrate this. The intuition in favour of incremental processing of dialogue contributions gets further support from studies of nonverbal behaviour in dialogue. Humans start to perform certain body movements that are perceived and interpreted by others as dialogue acts while the partner is still speaking. All this speaks strongly in favour of the incremental interpretation of dialogue behaviour.

Psycholinguistic studies provide further support for this view. For example, eye-tracking experiments reported by Tanenhaus et al. (1995), Sedivy et al. (1999) and Sedivy (2003) show that definite descriptions are resolved incrementally when the referent is visually accessible. Other evidence suggests that understanding involves *parallel* generation of multiple hypothesis. It has been shown, e.g. for processing ambiguous words by Swinney (1979) and Simpson (1994), for definite expression resolution (Tanenhaus et al., 1995), and for pronoun interpreta-

[1] The SCHISMA corpus consists of 64 dialogues in Dutch collected in Wizard-of-Oz experiments, has keyboard-entered utterances within the information exchange and transaction task domain, where users are supposed to make inquiries about theatre performances scheduled and make ticket reservations.

tion (Corbett and Chang, 1983), that all possible hypotheses are activated in parallel until it is possible to identify a single candidate or reduce their number.

The experimental evaluation of a non-incremental dialogue system and its incremental counterpart, reported in (Aist et al., 2007), showed that the latter is faster overall due to the incorporation of pragmatic information at early stages of the understanding process. Since users formulate utterances incrementally, partial utterances may be available for a substantial amount of time and may be interpreted by the system. An incremental interpretation strategy may allow the system to respond more quickly, by minimizing the delay between the time the user finishes and the time the utterance is interpreted (DeVault and Stone, 2003). For instance, the language understanding module typically performs the following tasks:

1. segmentation: identification of relevant segments in the input, such as sentences;

2. lexical analysis: lexical lookup, possibly supported by morphological processing, and by additional resources such as WordNet, VerbNet, or lexical ontologies;

3. parsing: construction of syntactic interpretations;

4. semantic analysis: computation of propositional, referential, or action-related content;

5. pragmatic analysis: determination of speaker intentions.

Of these tasks, lexical analysis, being concerned with local information at word level, can be done for each word as soon as it has been recognized, and is naturally performed as an incremental part of utterance processing, but syntactic, semantic and pragmatic analysis are traditionally performed on complete utterances. Tomita's pioneering work in left-to-right syntactic parsing has shown that incremental parsing can be much more efficient and of equal quality as the parsing of complete utterances (Tomita, 1986). Computational approaches to incremental semantic and pragmatic interpretation have been less successful (see e.g. Haddock, 1989; Milward and Cooper, 2009), but work in computational semantics on the design of underspecified representation formalisms has shown that such formalisms, developed originally for the underspecified representation of quantifier scopes, can also be applied in situations where incomplete input information is available (see e.g. Bos, 2002; Bunt, 2007a, Hobbs, 1985b, Pinkal, 1999) and as such hold a promise for incremental semantic interpretation.

Although language processing is largely incremental, some decisions need to be postponed. In some cases, a hypothesis cannot be resolved immediately, because there is insufficient evidence for disambiguation. Some semantic phenomena that cannot be resolved incrementally, e.g. scope assignment; here, partial interpretations may initially be constructed and refined later. We will show that in order to arrive at the best output prediction two different classification strategies are needed: (1) local classification that is based on features observed in dialogue behaviour and that can be extracted from the annotated data; and (2) global classification that takes the locally predicted context into account.

In this chapter we present the results of a series of experiments carried out in order to assess the automatic incremental segmentation and classification of dialogue acts. We investigate the automatic recognisability of multiple communicative functions on the basis of the observable features such as linguistic cues, intonation properties and dialogue history.

7.1 Classification experiments

7.1.1 Data and features

In our recognition experiments we used data selected from the AMI meeting corpus and MapTask dialogues, described in Section 4.5.1. For training we used three annotated AMI meetings that contain 17,335 tokens which form 3,897 functional segments. The MapTask training set contains 6 dialogues consisting of 5,941 tokens that form 2,589 functional segments. Only independent communicative functions were considered in the recognition experiments. Table 7.1 shows the distribution of annotated dialogue acts that belong to a particular dimension for both corpora by indicating the percentage of identified functional segments per dimension. Table 7.2 presents the percentage of functional segments with general-purpose functions. Note that for better recognition of pragmatic and semantic distinctions between different types of Inform acts they are divided into two categories: Informs and Informs that are rhetorically related to previous dialogue acts, e.g. elaborate, justify or explain them.

Table 7.1: Distribution of functional segments across dimensions for the AMI and MapTask corpora.

Dimension	AMI corpus	MapTask corpus
Task	31.8	52.4
Auto-Feedback	20.5	15.7
Allo-Feedback	0.7	4.7
Turn Management	50.2	24.3
Social Obligation Management	0.5	0.1
Discourse Structuring	2.8	0.5
Own Communication Management	10.3	2.8
Time Management	26.7	13.4
Partner Communication Management	0.3	0.3
Contact Management	0.1	1.7

The features included in the data sets considered here are those relating to *dialogue history*, *prosody*, and *word occurrence*.

For dialogue history we used of the tags of the 10 previous turns. Additionally, the tags of utterances to which the utterance in focus was a response, as well as timing: token *duration* and *floor-transfer offset*[2] computed in milliseconds, are included as features. For the segmented data (segmented per dimension), the occurrence of a segment inside another segment is encoded as a feature.

Prosodic, durational and temporal features were described in Section 6.1.

We also include the speaker (A, B, C, D) and the addressee (other participants individually or the group as a whole) as features.

For the classification experiments based on complete segments, word occurrence is represented by a bag-of-words vector[3] indicating the presence or absence of words in the segment. In total, 1,668 features are used for the AMI data and 829 features for the MapTask data.

In the learning experiments for incremental segment processing each token is coded as a feature. Additionally, bi- and trigram models were constructed and used as lexical features.

[2]Difference between the time that a turn starts and the moment the previous turn ends.

[3]With a size of 1,640 entries for AMI data and 802 entries for the MapTask data

Table 7.2: Distribution of functional tags for general-purpose communicative functions for the AMI and MapTask corpora (in %).

General-purpose function	AMI corpus	Maptask corpus
PropositionalQuestion	5.8	7.1
Set Question	2.3	2.9
Check Question	3.3	7.1
Propositional Answer	9.8	4.3
Set Answer	3.9	2.4
Inform	11.7	7.8
Inform (Rhetorical)	21.9	13.4
Instruct	0.3	26.8
Suggest	10.1	0

7.1.2 Classifiers

Of the many machine-learning techniques which have been used for NLP tasks, it is still an open issue which techniques are best suited for which task. We used three different types of classifiers to test their performance on the dialogue data: a probabilistic one, a rule inducer and a memory-based learner.

As a probabilistic classifier we used *Bayes Nets*. This classifier estimates probabilities rather than produce predictions, which is often more useful, because this allows us to rank predictions. Bayes Nets estimate the conditional probability distribution on the values of the class attribute given the values of the other attributes:

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$$

Bayes' Theorem expresses that the *posterior* probability distribution over a hypothesis *h*, given available data *D*, is to be computed from the *likehood* $P(D|h)$, i.e. the probability distribution of the data given a hypothesis, and the *prior* P(h), i.e. the prior probability distribution over possible hypotheses. We used the K2 search algorithm (Cooper and Herskovits, 1992) run with random number of orderings. Conditional probability distributions are estimated by using maximum likehood estimation, i.e. choose the hypothesis that maximises the likehood:

$$h_{ML} = argmax_h P(D|h)$$

Bayesian classifiers often work quite well for complex real-world situations and are particularly suitable for situations in which the dimensionality of the input is high. Moreover, this classifier requires relatively little computation, can be efficiently trained and represents conditional distribution in an easily comprehensible form.

As a rule induction algorithm, we chose *Ripper* (Cohen, 1995). The advantage of such an algorithm is that the regularities discovered in the data are represented as human-readable rules.

The third classifier is *IB1*, which is a memory-based learner that is a successor of the *k*-nearest neighbour (*k*-NN) classifier. The algorithm first stores a representation of all training examples in memory. When classifying new instances, it searches for the *k* most similar examples (nearest neighbours) in memory according to a similarity metric, and extrapolates the target class from this set to the new instances. The algorithm may yield more precise results

Table 7.3: Matrix for different outcomes of a two-class prediction.

		Predicted class	
		yes	no
Actual class	yes	true positive (tp)	false negative (fn)
	no	false positive (fp)	true negative (tn)

given sufficient training data, because it does not abstract away low-frequent phenomena during the learning (Daelemans et al., 1999).

The results of all experiments were obtained using 10-fold cross-validation[4].

In view of the relatively low frequencies of the tags in some dimensions, we use a baseline that is based on a single feature, namely the tag of the previous dialogue utterance (see Lendvai et al., 2003), unless specified differently.

7.1.3 Evaluation metrics

Several metrics have been proposed in the literature for the evaluation of classifier performance.

For assessing the performance of the joint segmentation and classification of dialogue acts, a word-based and a dialogue act-based metric are used. The word-based strict metric has been introduced in (Ang et al., 2005). It measures the percentage of words that were placed in a segment perfectly identical to that in the reference. In other words, if an output segment perfectly matches a corresponding reference segment on the word level, each word in that segment is counted as correct. All other words are counted as incorrect. A dialogue act-based metric (DER) was proposed in (Zimmermann et al., 2005). For the strict metric, a word is considered to be correctly classified if and only if it has been assigned the correct dialogue act type and it lies in exactly the same segment as the corresponding word of the reference. Thus, the DER metric not only requires a dialogue act candidate to have exactly matching boundaries but also to be tagged with the correct dialogue act type. We use the combined DER_{sc} metric to evaluate joint segmentation (s) and classification (c):

$$DER_{sc} = \frac{Tokens\ with\ wrong\ boundaries\ and/or\ function\ class}{total\ number\ of\ tokens} \times 100$$

The most commonly used performance metrics are accuracy, precision, recall and F-scores (harmonic mean). The overall success rate (*accuracy*) is computed by dividing the number of correct classifications by the total number of classifications. The proportion of correctly classified positive instances from all classified positive instances is known as *precision*, and the proportion of correctly classified instances from all positive instances is known as *recall*.

A common metric which represents the balance between precision and recall is the *F-score*:

$$F - score = \frac{2 \cdot recall \cdot precision}{recall + precision} = \frac{2 \cdot tp}{2 \cdot tp + fp + fn}$$

[4]In order to reduce the effect of imbalances in the data, it is partitioned ten times. Each time a different 10% of the data is used as test set and the remaining 90% as training set. The procedure is repeated ten times so that in the end, every instance has been used exactly once for testing (Witten and Frank, 2000) and the scores are averaged. The cross-validation was stratified, i.e. the 10 folds contained approximately the same proportions of instances with relevant tags as in the entire dataset.

We will use these standard metrics when evaluating classification results.

7.1.4 Incremental dialogue act classification

As we pointed out in the introduction to this chapter, human language processing is incremental, and dialogue system performance will benefit from the incorporation of pragmatic information at early stages of the understanding process. In this section we test the performance of classifiers based on partial input that is available for interpretation.

7.1.5 Related work

Nakano et al. (1999) proposed a method for incremental understanding of user utterances whose boundaries are not known. The *Incremental Sentence Sequence Search* (ISSS) algorithm finds plausible boundaries of utterances, called significant utterance (SU), that can be a full sentence or a subsentential phrase, such as a noun phrase or a verb phrase. Any phrase that can change the belief state is defined as SU. In this sense an SU corresponds more or less with what we call a functional segment. ISSS maintains multiple possible belief states, and updates those belief states when a word hypothesis is input (i.e. word-by-word). The ISSS approach does not deal with the multifunctionality of segments, however, and does not allow segments to overlap.

Lendvai and Geertzen (2007) proposed *token-based* dialogue act segmentation and classification, worked out in more detail in (Geertzen, 2009). This approach takes dialogue data that is not segmented into syntactically or semantically complete units, but operates on the transcribed speech as a stream of words and other vocal signs (e.g. laughs or breathing), including disfluent elements (e.g. abandoned or interrupted words) for each dialogue participant. Segmentation and classification of dialogue acts are performed simultaneously in one step. Geertzen (2009) reports on classifier performance on this task for the DIAMOND data using DIT^{++} labels; F-scores range from 47.7 to 81.7. It was shown that performing segmentation and classification together results in better segmentation performance, but affects the dialogue act classification negatively.

The incremental dialogue act recognition method as proposed here takes the token-based approach for building classifiers for the recognition of multiple dialogue acts for each input token, and adopts the ISSS idea of word-by-word information-state updates. Assigning priorities for potential updates we suggest to use posterior probability corresponding to constituent token in a given context estimated in the training experiments combined with empirical, pragmatic and logical constraints on dialogue act combination that will be discussed in detail in Section 8.4.1.

7.1.6 Classification results

We performed token-based machine-learning experiments on the AMI and MapTask data. The functional segment boundaries were encoded as follows: to each token its communicative function label was assigned as well as whether it starts a segment (B), is inside a segment (I), ends a segment (E), is outside a segment (O), or forms a functional segment on its own (BE). Thus, the class labels in the training data consist of segmentation prefixes (IBOE) and communicative function labels as shown in Table 7.6.

The results for joint segmentation and classification for different classifiers are presented in Table 7.4 for the AMI data.

Table 7.4: Overview of F-scores and DER for joint segmentation and classification in each DIT^{++} dimension for AMI data.

Classification task	BL		BayesNet		Ripper	
	F_1	DER	F_1	DER	F_1	DER
Task management	32.7	51.2	52.1	48.7	**66.7**	42.6
Auto-Feedback	43.2	84.4	**62.7**	33.9	60.1	45.6
Allo-Feedback	70.2	59.5	**73.7**	35.1	71.3	49.1
Turn Management: initial	34.2	95.2	**57.0**	58.4	54.3	81.3
Turn Management: final	33.3	92.7	**54.2**	46.9	49.3	87.3
Time management	43.7	96.5	**64.5**	46.1	61.4	53.1
Discourse Structuring	41.2	35.1	**72.7**	19.9	50.2	30.9
Contact Management	59.9	53.2	71.4	49.9	**83.3**	37.2
OCM	36.5	87.9	**68.3**	51.3	58.3	76.8
PCM	49.5	59.0	**58.5**	45.5	51.4	58.7
SOM	34.5	47.5	**86.5**	35.9	83.3	44.3

The results show that both classifiers outperform the baseline by a broad margin. The BayesNet classifier marginally outperforms the Ripper rule-inducer, showing no significant differences in overall performance. Comparing our results with those reported in (Geertzen, 2009) for the DIAMOND data, we see that the F-scores obtained in our experiments are slightly higher. This may be due to the fact that our training set is three times larger. For better comparison, we decided to perform the same experiments using MapTask dialogues. Table 7.5 shows the overall performance of the classifiers for joint segmentation and classification task for these data.

Table 7.5: Overview of F-scores and DER for joint segmentation and classification in each DIT^{++} dimension for MapTask data.

Classification task	BL		BayesNet		Ripper	
	F_1	DER	F_1	DER	F_1	DER
Task management	43.8	70.2	**79.7**	41.9	77.7	58.5
Auto-Feedback	64.6	60.6	65.4	55.2	**80.1**	43.9
Allo-Feedback	30.7	91.2	59.3	54.0	**72.7**	51.8
Turn Management	50.3	47.5	70.8	40.9	**81.4**	36.2
Time management	54.2	28.4	72.1	20.3	**83.6**	10.4
Discourse Structuring	33.2	95.1	62.5	44.3	**66.7**	43.5
Contact Management	24.7	93.2	**57.0**	79.5	11.0	93.5
OCM	11.2	97.4	**42.9**	64.7	28.6	92.1
PCM	14.3	95.2	61.5	55.2	**66.7**	50.1
SOM	08.8	96.2	40.0	71.8	**85.7**	21.4

Table 7.6: Segment boundaries and communicative function encoding.

S	Token	Task	Auto-F.	Allo-F.	TurnM.	TimeM.	ContactM.	DS	OCM	PCM	SOM
B	it	B:inf	O	O	O	O	O	O	O	O	O
B	has	I:inf	O	O	O	O	O	O	O	O	O
B	to	I:inf	O	O	O	O	O	O	O	O	O
B	look	I:inf	O	O	O	O	O	O	O	O	O
B	you	O	O	B:check	O	O	O	O	O	O	O
B	know	O	O	E:check	O	O	O	O	O	O	O
B	cool	I:inf	O	O	O	O	O	O	O	O	O
D	mmhmm	O	BE:positive	O	O	O	O	O	O	O	O
B	and	I:inf	O	O	BE:t_keep	O	O	O	O	O	O
B	gimmicky	E:inf	O	O	O	O	O	O	O	O	O

The classifiers' performance on the MapTask data is better than those on the AMI data, for the five most frequently occurring dimensions: Task, Auto- and Allo-Feedback, Turn Management and Time Management. This is because the classifiers need to deal with less complex phenomena and mechanisms in these data.

Although the results are encouraging, the performance for joint segmentation and classification does not outperform the two-step segmentation and classification scores reported in (Geertzen et al., 2007) and summarized in Table 7.7. It was noticed that lower F-scores are due lower recall. Beginnings and endings of segments were often not found. For example, the beginnings of Set Questions are identified with perfect precision (100%), but about 60% of the cases were not found. The reason that classifiers still show reasonable performance is that most tokens occur inside segments and were better classified, e.g. inside-tokens of Set Questions were classified with high precision (83%) and reasonably high recall scores (76%). In general, the correct identification of the start of a relevant segment is crucial for further decisions. This led us to the conclusion that the search space and the number of initially generated hypotheses for classifiers should be reduced, e.g. by splitting up the learning task which makes it more manageable. A widely used strategy is to split a multi-class learning task into several binary learning tasks. Learning multiple classes, however, allows a learning algorithm to exploit interactions among classes. We split the classification task in such a way that a classifier needs to learn (1) communicative functions in isolation; (2) semantically related functions together, e.g. all information-seeking functions (questions) or all information-providing functions (all answers and all informs).

Table 7.7: Overview of F-scores on the baseline (BL) and the classifiers on two-step segmentation and classification tasks.

Classification task	BL	NBayes	Ripper	IB1
Task	66.8	71.2	**72.3**	53.6
Auto-Feedback	77.9	86.0	**89.7**	85.9
Allo-Feedback	79.7	**99.3**	99.2	98.8
Turn M.: initial	93.2	92.9	93.2	88.0
Turn M.: final	58.9	85.1	**91.1**	69.6
Time management	69.7	99.2	**99.4**	99.5
Discourse Structuring	69.3	**99.3**	**99.3**	99.1
Contact Management	89.8	99.8	99.8	99.8
OCM	89.6	90.0	**94.1**	85.6
PCM	99.7	99.7	99.7	99.7
SOM	99.6	99.6	99.6	99.6

Classification of general-purpose functions

We present the segmentation and classification (as a joint task) results for each type of general-purpose function, and show that more accurate predictions are obtained when dealing with one particular kind of information. Both communicative function recognition and detection of segment boundaries improve significantly.

Segments having a general-purpose function may address any of the ten dimensions. We will argue below that recognition of the dimension should not be performed simultaneously with communicative function recognition, because it is too early at the beginning of a segment to say with reasonable certainty what type of semantic content is addressed. This decision is

Table 7.8: Overview of F-scores and DER for joint segmentation and classification for information-seeking communicative functions in AMI and MapTask data.

Classification task	BL		BayesNet		Ripper	
	F_1	*DER*	F_1	*DER*	F_1	*DER*
AMI data						
Propositional Question	47.0	39.1	**94.9**	3.9	75.8	23.5
Check Questions	43.8	56.4	**68.5**	19.6	61.3	33.1
Set Questions	44.8	52.1	74.1	18.6	**76.3**	17.7
Choice Question	41.8	54.2	68.6	15.7	**73.1**	21.4
MapTask data						
Propositional Question	29.5	73.9	**87.8**	13.5	71.6	27.1
Check Questions	25.0	73.2	**59.8**	63.6	52.8	57.9
Set Questions	24.8	72.6	**69.3**	42.2	69.0	43.1
Choice Question	23.5	73.4	**66.7**	48.9	67.1	45.7

better postponed until enough evidence from the speaker's behaviour is obtained. The question at what stage of processing a segment such a decision can be made, is addressed below.

Classifying information-seeking functions

In DIT^{++} four information-seeking functions are defined corresponding to four types of questions: Propositional Question, Check Question, Set Question and Choice Question.

We trained the classifiers to identify and classify questions; Table 7.8 gives an overview of the success scores. Both the recognition of questions and that of segment boundaries is fairly accurate.

Table 7.9: Overview of F-scores and DER for joint segmentation and classification for information-providing functions in AMI and MapTask data.

Classification task	BL		BayesNet		Ripper	
	F_1	*DER*	F_1	*DER*	F_1	*DER*
AMI data						
Inform	45.8	39.9	**79.8**	18.7	66.5	30.5
Inform (Elaborate)	37.2	38.9	**69.1**	13.4	68.7	23.9
Inform (Justify)	46.3	35.2	**80.5**	11.2	75.7	31.6
Inform (Conclude)	43.2	48.5	**66.7**	13.5	59.0	37.2
Inform (Remind)	47.5	38.6	**63.3**	21.4	56.2	22.7
(Dis-)Agreement	41.3	79.1	**72.1**	12.6	71.6	60.2
PropositionalAnswer	32.0	77.8	**66.8**	26.1	52.2	53.8
(Dis-)Confirm	25.0	87.3	**47.3**	30.3	46.5	47.2
Set Answer	44.3	54.2	**77.5**	13.2	57.3	44.1
MapTask data						
Inform	24.1	72.7	**69.3**	50.9	59.8	60.7
Inform (Clarify)	24.8	73.7	**65.0**	46.7	60.5	54.8
Inform (Elaborate/Explain)	16.3	71.7	47.8	62.7	**62.2**	60.9
Propositional Answer	19.6	70.7	63.3	58.2	**76.0**	41.7
Set Answer	24.8	73.0	61.5	38.9	**63.8**	40.6

Table 7.10: Overview of F-scores and DER for joint segmentation and classification for action-discussion functions in AMI and MapTask data.

Classification task	BL		BayesNet		Ripper	
	F_1	DER	F_1	DER	F_1	DER
AMI data						
Suggest	45.8	38.4	**65.6**	17.3	48.8	35.6
Request	45.8	49.3	**75.8**	14.5	50.3	36.9
Instruct	46.3	49.3	**60.5**	14.5	46.3	36.9
Address Request	34.8	74.8	**79.0**	15.3	54.2	42.1
Offer	25.0	93.7	**65.3**	23.9	45.6	34.3
MapTask data						
Instruct	36.0	66.3	**74.3**	26.7	69.5	41.3

Classifying information-providing functions

The category of information-providing functions includes several types of Inform functions, such as Agreement, Disagreement and Correction; and two main types of Answers: Propositional Answer and Set Answer. Confirm and Disconfirm are special cases of Propositional Answer and are provided as a reaction to Check Questions.

Table 7.9 gives an overview of the obtained F-scores for information-providing function classification. In general, classifiers performed well on this task. F-scores achieved are higher than baseline scores, ranging from 63.3 to 80.5 with one exception for (Dis-)Confirm acts, which are often confused with Propositional Answers because they share the same vocabulary. Since (Dis-)Confirm acts entail Propositional Question acts, this has little influence on the system's overall performance.

Classifying action-discussion functions

Request, Instruct, Suggestion and Accept/Reject Offer are defined as directives distinguished by the degree of pressure that the speaker puts on the addressee and the speaker's assumptions about the addressee's ability and agreement to perform a certain actions. Commissive acts such as Accept/Reject Request or Suggestion and Offer capture the speaker's commitments to perform certain actions.

Table 7.10 presents the classification results for action-discussion acts. Action-discussion acts are generally well recognized, the performance of BayesNets being better than that of Ripper.

Table 7.11: Overview of F-scores and DRE for complex label classification (boundary+communicative function+dimension) recognition in AMI data.

Classification task	BL		BayesNet		Ripper	
	F_1	DER	F_1	DER	F_1	DER
Task	28.0	83.2	**62.0**	78.2	48.0	77.8
Auto-Feedback	31.2	85.7	**45.3**	68.3	33.3	69.1
Allo-Feedback	23.3	96.2	**37.6**	80.3	24.7	83.6
Discourse Structuring: delimitation	24.5	93.9	**32.9**	87.1	29.4	91.6
Discourse Structuring: topic organisation	13.3	87.1	**23.0**	63.8	20.1	69.5

Dimension recognition

Dimension recognition can be approached in two ways. One approach is to learn segment boundaries, communicative function label and dimension in one step (e.g. the class label *B:task;inform*). This task is very complicated, however. First, it leads to data which are high dimensional and sparse, which will have a negative influence on the performance of the classifiers. Second, in many cases the dimension can be recognized reliably only with some delay; for the first few segment tokens it is often impossible to say what the segment is about. For example:

(79) 1. What do you think who we're aiming this at?
2. What do you think we are doing next?
3. What do you think Craig?

The three Set Questions in (79) start with exactly the same words, but they address different dimensions: question 1 is about the Task (in AMI - the design the television remote control); question 2 serves the purpose of Discourse Structuring; and question 3 elicits feedback.
Another approach is to first recognize segment boundaries and communicative function, and define dimension recognition as a separate classification task.

Table 7.12: Overview of F-scores for dimension recognition for general-purpose functions in AMI data.

Classification task	GP functions		Dimension recognition for GP functions			
	F_1	*DER*	Task	Auto-F.	Allo-F.	DS
Propositional Question	94.9	3.9	99.0	84.4	91.0	81.6
Set Question	74.1	18.6	94.8	79.6	na	87.5
Check Question	68.5	19.6	94.1	76.5	80.6	86.8
Inform	79.8	18.7	93.8	76.6	na	86.5
Inform (Elaborate)	69.1	13.4	94.6	na	58.3	86.9
Inform (Justify)	80.5	11.2	94.2	76.8	na	86.6
Inform (Conclude)	66.7	13.5	94.3	na	na	86.9
Inform (Remind)	63.3	21.4	94.1	na	na	86.9
(Dis-)Agreement	72.1	12.6	94.1	76.8	57.9	86.8
Propositional Answer	66.8	26.1	94.0	76.8	58.1	86.9
(Dis-)Confirm	47.3	30.3	94.1	76.7	58.0	na
Set Answer	77.5	13.2	94.1	76.7	na	86.8
Suggest	65.6	17.3	96.1	77.1	na	97.1
Request	75.8	14.5	99.1	na	na	86.8
Instruct	60.5	14.5	99.4	na	na	96.8
Address Request	79.0	15.3	99.0	na	na	86.2
Offer	65.3	23.9	96.1	na	na	78.3

Tokens	SetQuestion		Task		Auto-F.		TurnM.		Complex label (BIOE:D;CF)	
	label	*p*	label	*p*	label	*p*	label	*p*	label	*p*
what	B:setQ	0.85	O	0.71	O	1	O	0.68	O	0.933
you	I:setQ	1	task	0.985	O	1	B:give	0.64	O	0.869
guys	I:setQ	1	task	0.998	O	1	E:give	0.66	O	0.937
have	I:setQ	1	task	0.997	O	1	O	1	I:task;setQ	0.989
already	I:setQ	1	task	0.996	O	1	O	0.99	I:task;setQ	0.903
received	I:setQ	1	task	0.987	O	1	O	1	I:task;setQ	0.813
um	O	0.93	O	0.89	O	1	BE:keep	0.99	O	0.982
in	I:setQ	1	task	0.826	O	1	O	0.89	I:task;setQ	0.875
your	I:setQ	1	task	0.996	O	1	O	0.99	I:task;setQ	0.948
mails	E:setQ	0.99	task	0.987	O	1	O	1	E:task;setQ	0.948

Figure 7.1: Predictions with indication of confidence scores (highest *p* class probability selected) for each token assigned by five trained classifiers simultaneously.

We tested both strategies. The F-scores for the joint learning of complex class labels range from 23.0 (DER_{sc} = 68.3) to 45.3 (DER_{sc} = 63.8) (See Table 7.11). The results are reported only for those dimensions that are addressed in our data by general purpose functions in a substantial number of cases. We have no or only very few examples of general-purpose functions used for Turn, Time, Contact, Own/Partner Communication and Social Obligation Management.

For dimension recognition as a separate learning task the F-scores are significantly higher, ranging from 70.6 to 97.7 (see Table 7.12). The scores for joint segmentation and function recognition in the latter case are those listed in Tables 7.8, 7.9 and 7.10. Figure 7.1 gives an example of predictions made by five classifiers for the input *What you guys have already received um in your mails*. In fact hypotheses about the type of semantic content are generated for each token. The probability score for the first segment tokens are, however, lower than for the other tokens belonging to the same segment.

Classification of dimension-specific functions

The realisation of dimension-specific functions is highly conventional: dialogue participants use certain formulae to take or keep the turn, to open or close the dialogue, to move from one topic to another, to signal positive or negative feedback, and so on. Table 7.13 presents the results of dimension-specific function recognition, respectively. Both corpora do not have dimension-specific functions for the Task dimension, which is why this dimension is left out. The MapTask data does not have dimension-specific functions in the Allo-Feedback dimension.

Table 7.13: Overview of F-scores and DER for joint segmentation and classification for DIT^{++} dimension-specific functions in AMI and MapTask data.

Classification task	BL		BayesNet		Ripper	
	F_1	DER	F_1	DER	F_1	DER
AMI data						
Auto-Feedback	57.1	23.5	**78.8**	13.2	66.7	15.5
Allo-Feedback	89.3	4.4	**95.1**	2.9	94.3	3.9
Turn Management	24.8	21.9	**72.8**	7.4	46.3	10.7
Time management	68.3	32.3	82.4	13.7	**92.8**	11.4
Discourse Structuring	40.7	13.6	72.6	2.5	**74.5**	1.7
Contact Management	21.4	48.6	89.2	5.7	**92.3**	3.6
OCM	26.7	48.6	**78.0**	11.6	68.1	20.0
PCM	33.4	18.2	77.8	8.5	**88.9**	6.5
SOM	60.0	18.7	88.9	8.3	**90.1**	5.5
MapTask data						
Auto-Feedback	51.7	36.9	67.2	27.6	79.5	13.5
Turn Management	50.3	47.5	70.8	40.9	**81.4**	36.2
Time management	54.2	28.4	72.1	20.3	**83.6**	10.4
Discourse Structuring	65.3	17.2	92.3	10.4	93.2	8.9
Contact Management	33.3	34.6	54.6	26.2	70.5	12.2
OCM	11.2	97.4	**42.9**	64.7	28.6	92.1
PCM	14.3	95.2	61.5	55.2	**66.7**	50.1
SOM	08.8	96.2	40.0	71.8	**85.7**	21.4

7.2 Managing local classifiers

7.2.1 Global classification and global search

As was shown in the previous section, given a certain input we got all possible output predictions (hypothesis) from *local classifiers*. We have built in total 64 classifiers for dialogue act recognition for the AMI data and 43 classifiers for the MapTask data. The difference in the number of classifiers is due to the fact that there are fewer general-purpose functions in the MapTask dialogues (9 comparing to 16 in the AMI corpus). Some predictions made by local classifiers are false, but once a local classifier has made a decision it is never revisited. Humans, by contrast, may revise their previous decisions while interpreting utterances. It is therefore important to base a decision not only on local features of the input, but to take *outputs of all local classifiers* into account as well. For example, making use of the partial output predicted so far, i.e. of the history of previous predictions, and taking this as features into the next classification step, would help to discover and correct errors and make more accurate predictions. This is known as the 'recurrent sliding window strategy' (see Dietterich, 2002) when the true values for previous predictions are used as features. However, this suffers not only from the label bias problem when a classifier overestimates the importance of certain features, but also depicts an unrealistic situation, since this information is not available to a classifier in real time. The solution has been proposed by Van den Bosch (1997) as 'adaptive training', when the actual predicted output of previous processing steps are used as features.

We trained higher-level classifiers (often referred to as 'global') that have, along with features extracted locally from the input data as described above, the partial output predicted so far from all local classifiers. We used five previously predicted class labels, assuming that long distance dependencies may be important, and taking into account that the average length of a functional segment in our data is 4.4 tokens. Table 7.14 gives an overview of the results. We can observe an improvement of about 10-15% on average (see Tables 7.4 and 7.5). The classifiers still make some incorrect predictions, because the decision is sometimes based on incorrect previous predictions. An optimized global search strategy may lead to further improvements of these results.

Table 7.14: Overview for F-scores and DER_{sc} when global classifiers are used for AMI and MapTask data, based on added predictions of local classifiers for five previous tokens.

Classification task	AMI data				MapTask data			
	BayesNet		Ripper		BayesNet		Ripper	
	F_1	DER_{sc}	F_1	DER_{sc}	F_1	DER_{sc}	F_1	DER_{sc}
Task	65.3	14.9	**79.1**	21.8	81.6	17.8	**82.4**	14.1
Auto-Feedback	72.9	8.1	**77.8**	7.2	77.2	26.5	**81.3**	17.6
Allo-Feedback	67.7	10.9	**74.2**	9.5	68.3	35.4	**74.3**	20.6
Turn Management:initial	**72.2**	11.5	69.5	11.4	**82.9**	11.4	81.4	18.4
Turn Management:close	82.7	5.0	**83.0**	4.9	**72.9**	29.1	67.2	28.9
Time Management	70.0	3.0	**73.5**	2.1	**91.3**	8.7	75.8	19.3
Discourse Structuring	**72.3**	4.9	63.7	3.6	78.1	19.3	**81.3**	17.3
Contact Management	79.1	4.5	**84.3**	4.6	**79.5**	17.9	78.5	18.9
OCM	66.0	2.4	**68.3**	2.3	**80.4**	17.6	67.3	28.9
PCM	**63.2**	7.8	59.5	11.4	**72.7**	33.2	66.7	29.1
SOM	**88.4**	0.9	81.6	1.7	**95.7**	6.3	**95.7**	6.4

A strategy to optimize the use of output hypotheses is to perform a global search in the output space looking for best predictions. Our classifiers do not just predict the most likely class for an instance, but also generate a distribution of output classes. Class distributions can be seen as confidence scores of all predictions that led to a certain state. Our confidence models are based on token-level information given the dialogue left-context (i.e. dialogue history, wording of the previous and currently produced functional segment). This is particular useful for dialogue act recognition because the recognition of intentions should be based on the system's understanding of discourse and not just on the interpretation of an isolated utterance. Searching the (partial) output space for the best predictions is not always the best strategy, however, since the highest-ranking predictions are not always correct in a given context. A possible solution to this is to postpone the prediction until some (or all) future predictions have been made for the rest of the segment. For training, the classifier then uses not only previous predictions as additional features, but also some or all future predictions of local classifiers (till the end of the current segment or to the beginning of the next segment, depending on what is recognized). This forces the classifier to not immediately select the highest-ranking predictions, but to also consider lower-ranking predictions that could be better in the context of the rest of the sequence.

Table 7.15: Overview for F-scores and DER_{sc} when global classifiers are used for AMI and MapTask data, based on added predictions of local classifiers for five previous and five next tokens.

Classification task	AMI data				MapTask data			
	BayesNet		Ripper		BayesNet		Ripper	
	F_1	DER_{sc}	F_1	DER_{sc}	F_1	DER_{sc}	F_1	DER_{sc}
Task	82.6	9.5	**86.1**	8.3	**85.8**	12.2	80.8	9.1
Auto-Feedback	81.9	1.9	**95.1**	0.6	84.4	15.0	**93.0**	7.6
Allo-Feedback	**96.3**	0.6	95.7	0.5	**95.3**	4.6	94.6	6.9
Turn Management:initial	**85.7**	1.5	81.5	1.6	89.5	8.2	**91.0**	8.0
Turn Management:close	90.9	3.8	**91.2**	3.6	**82.9**	17.1	77.2	18.9
Time management	90.4	2.4	**93.4**	1.7	**94.9**	5.5	92.8	6.1
Discourse Structuring	**82.1**	1.7	78.3	1.8	85.7	12.4	**87.4**	8.2
Contact Management	87.9	1.2	**94.3**	0.6	87.4	9.9	**88.3**	7.4
OCM	78.4	2.2	**81.6**	2.0	87.2	9.8	**87.4**	7.6
PCM	**71.8**	2.4	70.0	4.6	86.7	11.1	**86.8**	9.8
SOM	98.6	0.4	98.6	0.5	**97.9**	1.1	**97.9**	1.2

Table 7.15 gives an overview of the global classification results based on added previous and next predictions of local classifiers. We can observe a further improvement in terms of high F-scores and quite low error rate. Both classifiers performed well on this task. The results show the importance of optimal global classification for finding the best output prediction. The use of local classifiers only is outperformed by a broad margin (see Tables 7.4 and 7.5 for AMI and MapTask data respectively). For instance, for one of most important dimension like Task F-scores reached by global classifiers is 86.1 while the best obtained F-scores training local classifiers is 79.7. F-scores for communicative function recognition by local classifiers range from 54.2 to 86.5, F-scores for communicative function recognition by global classifiers range from 71.8 to 97.9, statistically significant ($p < .05$, one-tailed z-test). Performance of global classifiers is very close to the performance of classifiers on two-step segmentation and classification task reported in Table 7.7. The performance of global classifiers is significantly

better on recognition of Task (F-scores of 86.1 compared to 72.3 for AMI data using Ripper and Auto-Feedback (95.1 compared to 89.7) acts that are the most important and frequently occurring dialogue acts in dialogue. It can also be concluded that the overall performance of global classifiers reported here is generally better than those reported in the literature (see Introduction to this chapter).

To summarize, we shown that token-based approach combining the use of local classifiers, which exploit local utterance features, and global classifiers which use the outputs of local classifiers applied to previous and subsequent tokens, is shown to result in excellent dialogue act recognition scores for unsegmented spoken dialogue. This can be seen as a significant step forward towards the development of fully incremental, on-line methods for computing the meaning of utterances in spoken dialogue.

7.3 Conclusions

This chapter presented a machine learning-based approach to the incremental understanding of dialogue utterances, with a focus on the recognition of their communicative functions. We discussed different strategies in the automatic recognition of dialogue acts. Not only word-level features are taken into account but also word N-grams, prosodic and acoustic features, and features calculated from speaker's and partner's previous utterances. The latter is particularly useful for communicative function recognition, because the recognition of a speaker's intention should be based on the system's understanding of the preceding discourse, and not just on understanding an utterance in isolation.

One of the main conclusions is that the commonly used strategy to first determine segment boundaries and subsequently perform dialogue act classification has serious theoretical and practical disadvantages. The identification of dialogue unit boundaries heavily depends on how a dialogue unit is defined (see Traum and Heeman, 1997). The definition of a functional segment is based on the criterion of carrying a communicative function: a functional segment is a minimal stretch of behaviour that has at least one communicative function. As a consequence, the identification of boundaries cannot precede the recognition of communicative functions.

Incremental dialogue act recognition is a complex task. Splitting up the output structure may make the task more manageable. Sometimes, however, learning of multiple classes allows a learning algorithm to exploit the interactions among classes. Combining these two strategies resulted in building a number of classifiers which show improvement both in communicative function recognition and in segment boundary detection, and result in excellent dialogue act recognition scores.

The incremental construction of input interpretation hypotheses has the effect that the understanding of an input segment is already nearly ready when the last token of the segment is received; viewing a dialogue act as a recipe for updating an information state, this means that the specification of the update operation is almost ready at that moment. It may even happen that the confidence score of a partially processed input segment is that high, that the system may decide to go forward and update its information state without waiting until the end of the segment, and prepare or produce a response based on that update. Of course, full incremental understanding of dialogue utterances includes not only the recognition of communicative functions, but also that of semantic content. However, many dialogue acts have no or only marginal semantic content, such as turn-taking acts, backchannels (*m-hm*) and other feedback acts (*okay*), time management acts (*Just a moment*), and in general dialogue acts with a dimension-specific function; for these acts the proposed strategy can work well without

semantic content analysis, and will increase the system's interactivity significantly. Moreover, given that the average length of a functional segment in our data is no more than 4.4 tokens, the semantic content of such a segment tends not to be very complex, and its construction therefore does not seem to require very sophisticated computational semantic methods, applied either in an incremental fashion (see e.g. Aist et al., 2007; DeVault and Stone, 2003) or to a segment as a whole.

Interactivity is however not the sole motivation for incremental interpretation. The integration of pragmatic information obtained from the dialogue act recognition module, as proposed here, at early processing stage can be beneficially used by the incremental semantic parser (but also syntactic parser module). For instance, information about the communicative function of the incoming segment at early processing stage can defuse a number of ambiguous interpretations, e.g. used for the resolution of many anaphoric expressions. A challenge for future work is to integrate the incremental recognition of communicative functions with incremental syntactic and semantic parsing, and to exploit the interaction of syntactic, semantic and pragmatic hypotheses in order to understand incoming dialogue segments incrementally in an optimally efficient manner.

Chapter 8

Context-driven dialogue act interpretation and generation

This chapter presents a context-based approach to the computational modelling of communicative behaviour in dialogue. We describe what constitutes a context in a dialogue and propose a context model that enables multiple simultaneous and independent updates in order to deal with the multifunctionality of dialogue contributions. We present the formal specification of updates on context models, and show how information is transferred from one dialogue participant to another. We discuss context update mechanisms and the communicative effects of the understanding of dialogue behaviour. These effects are the basis for dialogue participants to react in a certain way. We outline a context-driven approach to dialogue act generation that enables the construction of dialogue contributions that are multifunctional by design, and allows dialogue systems to apply a variety of dialogue strategies and communication styles.

Introduction

While the multifunctionality of dialogue utterances has been widely recognised, computationally oriented approaches to dialogue generally see multifunctionality as a problem, both for the development of annotation schemes and for the design of dialogue systems (Traum, 2000). As a consequence, information that may be obtained through a multifunctional analysis is often sacrificed for simplicity in computational modelling. Existing dialogue systems such as TRIPS (Allen et al., 2001), GoDIS (Larsson et al., 2000) or ViewGen (Wilks and Balim, 1991) gen-

Section 8.1 summarizes the work of the entire Dialogue Act Group at Tilburg University done during past 10 years, in particular in collaboration with the group leader Harry Bunt, Roser Morante, Simon Keizer and Jeroen Geertzen. Section 8.2 is inspired by Bunt's keynote speech at IWCS in Oxford, see Bunt (2011). Section 8.3 is base on Petukhova and Bunt (2010a) and got its current shape in the course of continuous discussions with Harry Bunt, also inspired by the work of Simon Keizer and Roser Morante. Parts of Section 8.4.1 are based on Petukhova et al. (2010), for which I did the research, in close collaboration with my co-authors. The software for automatic entailment and implicatures generation and checking has been designed by Andrei Malchanau. Section 8.4.2 is inspired by the work of Marcin Włodarczak (2009); experiments reported here were designed in close cooperation with Marcin, the interpretation of the results is mine. Section 8.4.3 is inspired by Keizer et al. (2011), I contributed to this book chapter as a co-author. The text of this section, however, is entirely mine. Section 8.4.4 presents original work.

erate an output by constructing an utterance which, when interpreted successfully by the user, satisfies a particular goal or step in a plan. Since a natural language utterance mostly has multiple communicative functions, however, the user may be expected to interpret the utterance as having all these functions, whereas the dialogue system produced the utterance in order to have just one of them. This is likely to lead to misunderstandings.

Some misunderstandings can be avoided by adequate computational modelling of the natural multifunctionality of spoken utterances. A crucial step in such a process is the construction of a context model that enables multiple updates. Context provides the basis for the interpretation of the speaker's behaviour and for decisions about future actions. An important issue is therefore what kinds of information should be included in a context model. In general, the term 'context' refers to the surroundings, circumstances, environment, background or settings of the activity of which the context is considered. In linguistics the term 'context' has most often been interpreted as referring to the surrounding text. Dialogue context is understood as the totality of conditions that influence the understanding and generation of communicative behaviour. This includes information about (a) the participants' information about the underlying task and its domain; (b) the participant's processing abilities and state of processing; (c) the availability and properties of communicative and perceptual channels and settings; (d) communicative obligations and constraints on the type of interaction; (e) participants' roles and social status in the dialogue; (f) information available to the dialogue participants before the dialogue and/or from the preceding dialogue contributions; (g) discourse plans. The dialogue context is partly dynamic, in the sense of changing during a dialogue as the result of the participants interpreting each other's communicative behaviour, reasoning with the outcomes of these processes, and planning further activities. These changes are essential in determining the continuation of the dialogue.

Most state-of-the-art dialogue systems contain a Dialogue Manager, a module which takes care of deciding which action to take next in the dialogue, given some form of information state or context model that is monitored and updated during the dialogue. A Dialogue Manager that generates dialogue acts from several dimensions simultaneously allows for less rigid system behaviour and therefore more natural interactions with users. A multidimensional approach opens the perspective of generating utterances which are multifunctional by design, rather than by accident. Keizer and Bunt (2007) showed that such a generation process allows dialogue systems to apply a variety of dialogue strategies and styles of communication. Issues such as whether or not to produce an explicit dialogue act that is already implied by another candidate dialogue act; what types of multifunctional utterances to generate in order to make the system act more efficiently; and which modalities to use for which functions, can be fruitfully investigated by using multiple dimensions in generation and evaluation, taking dependency relations between candidate dialogue acts into account.

The chapter is structured as follows. We present a multidimensional context model (Section 8.1), and discuss the update operators defined in DIT (Section 8.2). Section 8.3 explores a context-driven approach to the generation of multiple dialogue acts showing how multiple dialogue acts correspond to multiple context update operations on addressee's context model and discussing update mechanisms and communicative effects. In Section 8.4 we discuss the selection process of admissible dialogue act combinations and explore the possibility to apply different interactive styles and define different dialogue strategies. Finally, the observations and considerations put forward in this chapter are summarized in Section 8.5.

8.1 Context model

An Information State (IS) according to DIT is represented by a Context Model (CM) which contains all information considered relevant for interpreting dialogue utterances (in terms of dialogue acts) and for generating dialogue acts (leading to utterances). In order to formulate an update semantics for multifunctional dialogue segments, we need an articulate context model that enables multiple simultaneous and independent updates. An utterance, when understood by a dialogue participant as a dialogue act with a certain communicative function and semantic content, evokes certain changes in the participant's context model. These changes typically do not affect the entire context model, but only certain parts of it. Which part of a context model is affected by a dialogue act depends on the type of its semantic content. Thus, a communicative function specifies how an understanding dialogue participant's context model is updated, where the dimension (semantic content type) determines which parts of the model are updated. Since DIT distinguishes 10 orthogonal dimensions, as described in previous chapters, it may seem plausible to have 10 context types. This is, however, not really necessary, since some types of information are closely related. For instance, time management is closely related to the processing state, and these two types can be combined in one context part. Own and Partner Communication management are both concerned with a participant's state of processing, of the speaker and of the addressee respectively. They can conveniently be treated as one context type.

$$
\left[\begin{array}{l}
FunctionalSegment(FS): \\
\end{array}\right.
\left[\begin{array}{l}
start : \langle token_{index} | time\_point \rangle \\
end : \langle token_{index} | time\_point \rangle \\
verbatim : \langle token_{index} | time\_points = \text{`}token1', \ldots \rangle \\
prosody : \langle duration, pitch, energy, \ldots \rangle \\
nonverbal : \langle \left[\begin{array}{l} head\langle element_{index} | time\_points = expression1, \ldots \rangle \\ hands\langle element_{index} | time\_points = expression1, \ldots \rangle \\ face\langle element_{index} | time\_points = expression1, \ldots \rangle \\ posture\langle element_{index} | time\_points = expression1, \ldots \rangle \end{array}\right] \rangle \\
sender : \langle participant \rangle \\
dial\_acts(DAs) : \{\langle \left[\begin{array}{l} dimension(D) : \langle dim \rangle \\ comm\_function(CF) : \langle cf \rangle \\ sem\_content(SC) : \langle content \rangle \\ sender/speaker : \langle participant \rangle \\ addressee(-s) : \{\langle participant \rangle\} \\ func\_dependency : \left[\begin{array}{l} antecedent : \{\langle DA \rangle\} \end{array}\right] \\ fb\_dependency : \left[\begin{array}{l} antecedent : \{\langle FS \rangle\} \end{array}\right] \\ rhetorical\_relation : \left[\begin{array}{l} antecedent : \{\langle DA \rangle\} \\ type : \langle elaborate | \ldots \rangle \end{array}\right] \end{array}\right] \rangle\}
\end{array}\right]
$$

Figure 8.1: Feature structure representation of a functional segment.

Analysing and modelling the semantics of dialogue acts tells us what kinds of information should be included into a context model. This includes information concerned with (1) the participants information about the underlying task and its domain, as well as their beliefs about the dialogue partner's information of this kind (*semantic context*); (2) the participants state of processing (*cognitive context*); (3) the availability and properties of communicative and perceptual channels, and the partner's presence and attention (*physical/perceptual context*); (4) communicative obligations and constraints (*social context*); (5) the preceding dialogue contributions and possible discourse plans (*linguistic context*) (see Bunt, 1994).

In combination with additional general conceptual considerations, the context model has evolved into a five-component structure:

1. **Linguistic Context (LC)**: information about functional segments (1) produced up to this point ('dialogue history'); (2) most recently produced segment ('latest state'); and (3) planned future contributions ('dialogue future' or 'planned state'). For each functional segment information about wording, prosodic and nonverbal properties is specified. Each element of this representation has an indication of either the number (index) referring to the source information (e.g. tokenized speech transcription) or time slot (start and end) within the given element is produced . A representation of a functional segment contains information about a sender and a dialogue act performed. The latter includes information about the speaker and the addressee(-s) , representation of the semantic content and pragmatic analysis such as type of semantic content or dimension, communicative function, referents/antecedents in the dialogue history, i.e. what previous functional segments (feedback dependency relation) or dialogue act (functional dependency relation) the performed act is referring to, as well as rhetorical relation to previous dialogue contributions with specification of type of such relations. See Figure 8.1.
2. **Semantic Context (SemC)**: information about the task that includes representation of (1) task progress and success; (2) speaker's beliefs about the domain ('domain knowledge'); (3) speaker's beliefs about the dialogue partner's semantic context[1].
3. **Cognitive Context (CC)**: information about (1) the current processing state of the speaker; (2) assumptions and expectations about the partner's cognitive context; (3) and estimation of time need for processing of the current contribution.
4. **Perceptual/Physical Context (PC)**: information about the perceptible aspects of the communication process and the task/domain such as speaker's presence and readiness to continue the dialogue and assumptions about partner's perceptual/physical context.
5. **Social Context (SocC)**: information about current speaker's (1) interactive pressures and (2) reactive pressures, and assumptions and expectations about partner's social context.

Each of these five components contains the representation of three parts: (1) the speaker's beliefs about the task, about the processing of previous utterances, or about certain aspects of the interactive situation; (2) the addressee's beliefs of the same kind, according to the speaker; and (3) the beliefs of the same kind which the speaker assumes to be shared (or 'grounded') with the addressee. Note that part (2) introduces full recursion in each component of the context model; it depends on the kind of application whether this is indeed necessary (see Bunt, 2000). Figure 8.2 illustrates the proposed context model with its component structure.

Each of the parts of the model can be updated while other parts may remain unaffected. For example, a question about the task domain and an answer to it trigger the updates in the context model illustrated in Table 8.1.

A participant asks a question because (i) he wants to know something; and (ii) he assumes that the addressee might posses this knowledge. Questions have an additional default function of assigning the turn to the addressee (or of releasing the turn when addressing a group of addressees), allowing the question to be answered. Under "normal input-output" conditions (Searle, 1969), i.e. where participants speak the same language, use communication channels without severe distortions, have no hearing or speaking disorder, and so on, speakers normally expect to be perceived, understood and accepted unless there is evidence to the contrary, which can be formally represented by a *weak belief* that the addressee of his dialogue act believes that

[1]The context model as described here is suggested for two-party dialogues. A context model for multiparty dialogues would be more complex containing representation of speaker's beliefs about contexts of one or more addressees' and possibly also of other participants, e.g. side-participants, overhearers, etc.

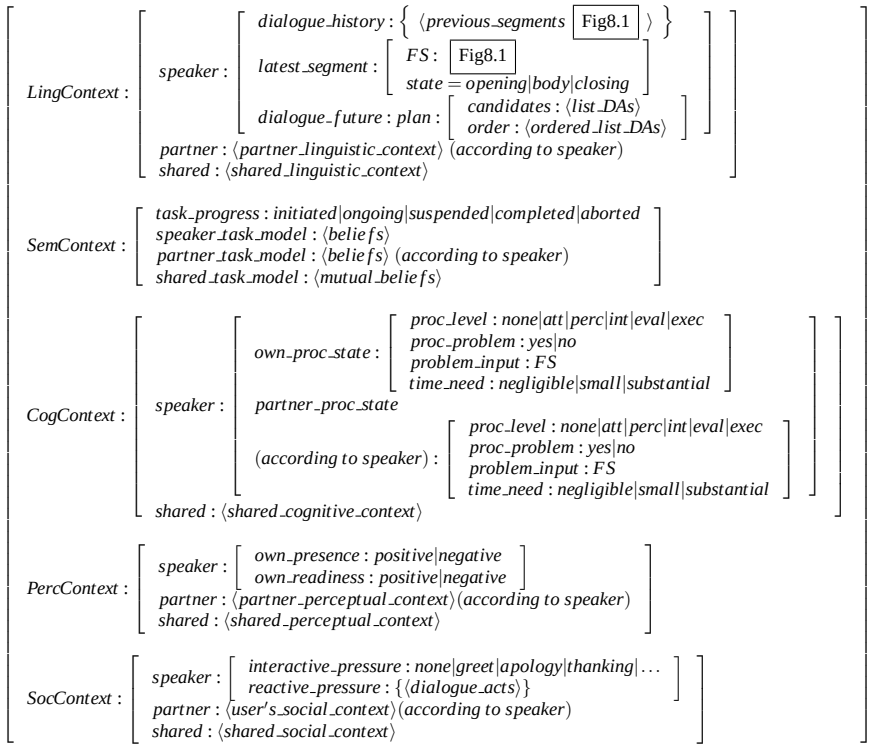


Figure 8.2: Feature structure representation of the context model.

the preconditions of the dialogue act are true (represented in his Semantic Context); that the addressee understands what is being said (represented in this stage in his Cognitive Context), and that the addressee believes that the turn is available for him (represented in his Linguistic Context). The assumptions of being understood and believed are not idiosyncratic for a particular speaker, but are commonly made by dialogue participants in cooperative dialogue under normal input-output conditions (see Morante (2007) and Bunt et al. (2007)). The condition S believes that U weakly believes that S believes that *p* leads to the conclusion that both S and U believe that it is *mutually believed* that U weakly believes that S believes that *p*. Thus, all dialogue participants have *mutual weak beliefs* about the three types of information listed above and expressed in s1,s2, and u1, u2; s5 and u3; and s7 and u4 in Table 8.1. If the addressee S understands the question, and if he actually adopts U's goal, then he tries to provide an answer. Preconditions for performing an answer are that S possesses the information in question and wants that U obtains this information. Since he believes that the turn is available for him, he is able to provide this information. S thus plans a dialogue act with the communicative function Answer and the requested information as its semantic content.

Table 8.1: Example of context update for Task Question-Answer pair. (LC = Linguistic Context; SC = Semantic Context; CC = Cognitive Context; prec = preconditions; du = dialogue utterance; da = dialogue act; fs = functional segment; D = dimension; CF = communicative function; exp.und = expected understanding; und = understanding; exp.ad = expected adoption; ad = adoption; Bel = believes; MBel = mutually believed; WBel = weakly believes)

Context	num	source/ *role*	S's context	num	source/ *role*	U's context
SC				u01 u02	*prec*	$Want(U, KnowVal(U,p))$ $Assume(U, KnowVal(S,p))$
LC				u03	*prec*	$Bel(U, Next\_Speaker(U))$
	s1		$Bel(S, Current\_Speaker(U))$	u1		$Bel(U, Current\_Speaker(U))$
	$fs_1 : du1$	*latest*	$\langle verbatim \rangle$	$fs_1 : du1$	*latest*	$\langle verbatim \rangle$
	$fs_1 : da_1$	*D;CF* *sem_content*	Task; Question $\langle content \rangle$ Speaker:U; Addressee:S	$fs_1 : da_1$	*D;CF* *sem_content*	Task; Question $\langle content \rangle$ Speaker:U; Addressee:S
	$fs_1 : da_2$	*default*	Turn-M.;Turn-Assign Speaker:U; Addressee:S	$fs_1 : da_2$	*default*	Turn-M.;TurnAssign Speaker:U; Addressee:S
	s2	exp.und: $fs_1 : da_2$	$Bel(S, MBel(\{S,U\}, WBel(U, Wants(U, Next\_Speaker(S))))$	u2	exp.und: $fs_1 : da_2$	$Bel(S, MBel(\{S,U\}, WBel(U, Next\_Speaker(S))))$
	s3	und:u2	$Bel(S, Want(U, Next\_Speaker(S))$			
	s4	ad:da_2	$Want(S, Next\_Speaker(S))$			
	s00	s4	$Want(S, Bel(U, Next\_Speaker(S)))$			
SC	s5a	exp.und: $fs_1 : da_1$	$Bel(S, MBel(\{S,U\}, WBel(U, Bel(S, Want(U, KnowVal(U,p)))$	u3	exp.und: $fs_1 : da_1$	$Bel(U, MBel(\{S,U\}, WBel(U, Bel(S, Want(U, KnowVal(U,p)))$
	s5b	exp.und: $fs_1 : da_1$	$Bel(S, MBel(\{S,U\}, WBel(U, Bel(S, Assume(U, KnowVal(S,p)))$	u4	exp.und: $fs_1 : da_1$	$Bel(U, MBel(\{S,U\}, WBel(U, Bel(S, Assume(U, KnowVal(S,p)))$
	s6a	und:u3	$Bel(S, Want(U, KnowVal(U,p)))$			
	s6b	und:u4	$Bel(S, Assume(U, KnowVal(S,p)))$			
	s7	ad:da1	$Want(S, KnowVal(U,p))$			
	s01	s8	$Bel(S,p))$			
CC	s8	exp.und: $fs_1 : du1$	$Bel(S, MBel(\{S,U\}, WBel(U, Interpreted(S, du1)))$	u4	exp.und: $fs_1 : du1$	$Bel(U, MBel(\{S,U\}, WBel(U, Interpreted(S, du1)))$
	s9	und:u4	$Bel(S, Interpreted(S, du1))$			
	s02	s9	$Want(S, Bel(U, Interpreted(S, du1)$			
LC	da_3	*plan:s00*	Turn Accept Speaker:S; Addressee:U antecedent: da_2			
	da_4	*plan:s02*	Auto-F.;Interpretation Speaker:S; Addressee:U antecedent: fs_1			
	da_5	*plan:s01*	Task; Answer Speaker:S; Addressee:U antecedent: da_1			

Update operations should however not undermine the consistency of the context model. Inconsistencies may occur for example when a dialogue participant changes his mind about information that has already been grounded. Consider the following example:[2]

(80) B1: Do you think then voice recognition is something we should really seriously consider?
C1: I thought we agreed not to include voice recognition

B's utterance is a Propositional Question, which corresponds to an update operation where the addressee C should add to his information about B that B wants to know whether to consider voice recognition. This is inconsistent with C's belief that it was decided earlier not to include voice recognition.

Before performing an update operation, an addressee should check whether the operation would keep his context model consistent. The DIT dialogue model which assumes several levels of processing by dialogue participants is useful for this purpose, where processing at the *evaluation* level checks whether update operations would keep the current context model consistent. If so, the updates are performed and the agent moves on to the stage of *execution* (for example, trying to compute the relevant answer information to the question). One way to implement this approach is to add to a context model a part called the *pending context*, which contains those changes that should be checked for maintaining consistency before they are applied. Thus, during the interpretation phase this information is stored in the pending context and it ends up in the main context model only after successful evaluation.

8.2 Update operators

Dialogue participants are assumed to be motivated, cooperative and rational agents. Motivations are captured in DIT by means of preconditions expressing a goal, which trigger the performance of a certain dialogue act. Additional conditions enable the performance of this dialogue act in a given context, e.g. conditions expressing beliefs about the current dialogue state (what has been understood and adopted), but also assumptions about the addressees' knowledge, expectations, and abilities.

When an addressee understands the speaker's behaviour this means that he is able to identify functional segments and their intended interpretation as dialogue acts. Understanding that a certain dialogue act is performed means believing that the preconditions which are characteristic for that dialogue act hold. These beliefs are thus added to an addressee's pending context.

Dialogue acts are often formally defined as operators that have certain effects on the speaker's and addressees' context models and are characterized in terms of *preconditions*, *effects* and a *body* that describes the means by which effects are achieved (see Allen, 1983). In DIT the update effects of a dialogue act derive from the specification of its preconditions as a consequence of general communication principles; this will be discussed in Section 8.3.

8.2.1 Semantic primitives

To specify the update semantics of communicative functions and describe the effects of dialogue acts on information states, we first need to specify some basic concepts ('semantic primitives') to represent the update effects. The update effects of dialogue acts relate directly

[2] From the AMI meeting corpus - ES2002b.

to their preconditions, since understanding an utterance as expressing a certain dialogue act A_i corresponds to believing that the speaker's context model satisfies the preconditions that are characteristic for A_i. In order to allow context models to support inferencing and planning, we cannot directly use the preconditions defining DIT dialogue acts, since these are expressed in natural language. The formalisation of these preconditions requires the formal representation of:

- beliefs about the task and about the dialogue;
- commitments, willingness and abilities to perform certain actions;
- goals that can be achieved by means of dialogue acts;
- assumptions about what addressees know and do not know; their goals; and their commitments, abilities, and willingness to perform certain actions.

The primitive concepts which are used to formulate the preconditions of the dialogue acts defined in DIT and the corresponding update effects are listed in the Tables 8.2 and 8.3 for dialogue acts with general-purpose and dimension-specific communicative functions, respectively.

Table 8.2: Semantic primitives for formalizing DIT dialogue act preconditions.

Concept	Definition
Mental attitudes towards information	
Bel(A,p,σ)	agent A possesses the information represented by the proposition p, with belief strength σ; σ represents A's certainty that p is true σ can have the values 'firm' and 'weak'
KnowVal(A,z)	A possesses information which allows him to compute the value of z
Want(A,τ)	agent A has goal τ
Mental attitudes towards actions	
CanDo(A,α)	agent A is able to perform action α
WillDo(A,α,C_α)	agent A is willing to perform action α if condition C_α is fulfilled; C_α may be 'empty', i.e. the universally true statement $\top$
CommitDo(A,α,C_α)	agent A is committed to perform action α if condition C_α is fulfilled; C_α may be 'empty', i.e. the universally true statement $\top$
CommitRefrain(A,α,C_α)	agent A is committed not to perform action α if condition C_α is fulfilled; C_α may be 'empty', i.e. the universally true statement $\top$
ConsidDo(A,α,B,C_α)	agent A is considering action α to be perform by B if condition C_α is fulfilled; C_α may be 'empty', i.e. the universally true statement $\top$; B may be identical to A
Interest(A,α)	means that action α is of interest to agent A

The first block of primitives in Table 8.2 serves to express an agent's attitudes towards information. The primitive **Bel** expresses the possession of information. It has three arguments: an agent whose beliefs are represented, a proposition which is believed, and the strength of the beliefs. The third argument may have two values: 'firm' and 'weak'[3]. For

[3]This is obviously a simplification compared to real situations. In reality, dialogue participants may have all sorts of beliefs, e.g. very certain or almost certain, or rather uncertain, or very uncertain. However, for the characterization of communicative function preconditions a binary distinction certain/uncertain suffices..

convenience the following shorthand notations are defined: **WBel**(S,p) = **Bel**(S,p,weak) and **Bel**(S,p) = **Bel**(S,p,firm). Covering both these cases, '**Assume**', is defined as **Assume**(A,p) iff **Bel**(A,p,firm) $\vee$ **Bel**(A,p,weak).

The primitive **KnowVal** serves to represent an agent's assumptions about the information available to another agent, and to represent information which an agent would like to possess. For example, an agent A believes that agent B knows whether *p* is true is represented as **Bel** (A, **KnowVal**(B,p), σ). The primitive **Want** is used to capture agent's goals to achieve a certain situation.

The second block of primitives serves to express an agent's attitudes towards actions. **CommitDo** and **CommitRefrain** are used to express an agent's commitment to perform and respectively not perform a certain action, possibly dependent on a condition. Such a condition may be specified in the semantic content of an utterance. **CanDo** and **WillDO** are used to express an agent's ability and willingness to perform a certain action. **ConsidDo** has two agent arguments: one (*A*) who considers an action α, and an other one (*B*) who would perform the action, where possibly $A = B$. **Interest**(A,α) expresses that action α is in the interest of agent *A*.

The formal and computational properties of operators expressing beliefs, goals and commitments have been studied extensively in logic and in Artificial Intelligence (see e.g. Moore, 1985; Konolidge, 1986, Hintikka, 1962, Cohen and Levesque, 1990, Fagin et al, 1995).

A third block of primitives, listed in Table 8.3, is used to characterize the effects of dialogue acts in other dimensions than Task. For example, the effects of feedback acts at different levels of processing can be represented using the predicates *Attended*, *Perceived*, *Interpreted*, *Evaluated* and *Executed*. Underspecified feedback acts of which the level of processing is not specified are interpreted as follows. Positive underspecified feedback relates the level of interpretation or higher; negative underspecified feedback relates to the level of interpretation or lower (Bunt, 2011). The context update effects of underspecified feedback acts can be represented by the predicates *Pos.Processed* and *Neg.Processed*, defined as $Pos.Processed = Executed \vee Evaluated \vee Interpreted$ and $Neg.Processed = \neg Attended \vee \neg Perceived \vee \neg Interpreted$.

The semantic primitives introduced here are rather similar to those proposed by Poesio and Traum (1998). Poesio and Traum's primitives are limited, however, to express update effects of a small set of 8 general-purpose functions and one positive Auto-Feedback function. The updates defined in DIT form a much wider variety of around 300 types of dialogue acts (see Bunt, 2011) and therefore require a more comprehensive set of primitives.

8.2.2 Update semantics of DIT communicative functions

The primitives defined above can be used to formally specify the update semantics of the dialogue acts defined in DIT. As already noted, a communicative function is interpreted as a function which, applied to a given speaker, addressee, and dimension, results in a function which can be applied to a semantic content in order to obtain a full context-update specification. Communicative functions in DIT are defined in terms of preconditions that trigger and enable the performance of a certain dialogue act. Each precondition corresponds to an update function (called *elementary update function*, see Bunt, 2011). Update functions when combined in a certain way are used to specify update effects of a dialogue act of a certain type. In this section the formalisation of preconditions of the DIT communicative functions are discussed.

Table 8.3: Semantic primitives for formalising the preconditions of DIT dialogue acts with dimension-specific communicative functions.

Concept	Definition
$Attended(A,s)$	agent A paid attention to the segment s
$Perceived(A,s)$	agent A successfully perceived the segment s
$Interpreted(A,s)$	agent A successfully interpreted the segment s
$Evaluated(A,s)$	agent A successfully evaluated the segment s
$Executed(A,s)$	agent A successfully executed the segment s
$Current\_Speaker(A)$	agent A currently occupies the speaker role
$Next\_Speaker(A)$	agent A is next to occupy the next speaker role
$Time\_Need(A, \langle small \vert substantial \rangle$	agent A needs a small or a substantial amount of time before proceeding with his contribution
$Present(A)$	agent A is present
$Ready(A)$	agent A is ready to receive and send messages
$Current-Topic(t_i)$	t_i is the topic addressed in the previous dialogue act
$Next-Topic(t_i)$	t_i is the topic addressed in the next dialogue act
$Mistake(A, fs_p)$	agent A made a mistake in the production of the segment part fs_p
$Delete(A, fs_p)$	agent A deletes segment part fs_p
$Substitute(A, (fs_p, fs_n))$	agent A substitutes the segment part fs_p for the segment part fs_n
$Concatenate(A, (fs_p, fs_n))$	agent A appends to the segment part fs_p the segment part fs_n
$Knows\_Id(A,B)$	agent A knows the identity of agent B
$Grateful(A,B,\mu)$	agent A is grateful for something μ performed by agent B, where μ is either information or an action
$Regret(A,\mu)$	agent A regrets that μ, where μ is either information agent A provided or an action agent A performed
$Final(A, fs)$	functional segment fs is the final segment contributed by A
$Open-dialogue$	propositional constant expressing that the dialogue is open
$Close-dialogue$	propositional constant expressing that the dialogue is closed

General-purpose functions

As seen in Chapter 4, the set of general-purpose communicative functions defined in DIT falls apart into information-transfer functions and action-discussion functions. Table 8.4 shows the formalised preconditions for the information-transfer functions, the upper part containing the information-seeking functions; the lower part the information-providing functions.

Action-discussion functions are divided into commissives and directives, expressing the speaker's commitment to perform a certain action and his wish that the addressee performs a certain action, respectively.

While accepting a request implies a commitment to perform the requested action, declining a request can be viewed as a commitment to *not* perform the requested action, and is therefore also a commissive act. Accepting and declining a request are two extremes on a scale of possible responses to a request. The function Address Request applies to those cases where the speaker responds to a request without committing himself to accepting or declining to perform the requested action, as in *Maybe later* in response to *Can you print this for me?*. Similarly for Address Offer and Address Suggest. Table 8.5 shows the formalised preconditions for the

Table 8.4: Formalised preconditions for the information-seeking and information-providing communicative functions defined in DIT (S = sender; A = addressee).

Communicative function	Preconditions
PropositionalQuestion	$Want(S, KnowVal(S, p))$ $Assume(S, KnowVal(A, p))$
CheckQuestion	$Want(S, KnowVal(S, p))$ $Assume(S, KnowVal(A, p))$ $WBel(S, p)$
SetQuestion	$Want(S, KnowVal(S, P))$ $Assume(S, KnowVal(A, P))$ $Bel(S, \exists x.P(x)))$
ChoiceQuestion	$Want(S, Bel(S, p) \vee Bel(S, q))$ $Assume(S, p \vee q)$ $Assume(S, Bel(A, p) \vee Bel(A, q))$
Inform	$Bel(S, p, \sigma)$ $Want(S, Bel(A, p, \sigma))$
Answer	$Bel(S, p, \sigma)$ $Want(S, Bel(A, p, \sigma))$ $Assume(S, Want(A, KnowVal(A, p)))$ $Assume(S, Assume(A, KnowVal(S, p)))$
Confirm	$Bel(S, p, \sigma)$ $Want(S, Bel(A, p, \sigma))$ $Assume(S, WBel(A, p))$ $Assume(S, Want(A, KnowVal(A, p)))$ $Assume(S, Assume(A, KnowVal(S, p)))$
Disconfirm	$Bel(S, \neg p, \sigma)$ $Want(S, Bel(A, \neg p, \sigma))$ $Assume(S, WBel(A, p))$ $Assume(S, Want(A, KnowVal(A, \neg p)))$ $Assume(S, Assume(A, KnowVal(S, p)))$
Agreement	$Bel(S, p, \sigma)$ $Want(S, Bel(A, p, \sigma))$ $Bel(S, Assume(A, p))$
Disagreement	$Bel(S, \neg p, \sigma)$ $Want(S, Bel(A, \neg p, \sigma))$ $Bel(S, Assume(A, p))$
Correction	$Bel(S, \neg p_1, \sigma)$ $Bel(S, p_2, \sigma)$ $Want(S, Bel(A, \neg p_1, \sigma))$ $Want(S, Bel(A, p_2, \sigma))$ $Assume(S, Assume(A, p_1))$

action-discussion functions in DIT.

Dimension-specific functions

Being a domain-independent taxonomy, DIT^{++} does not include dimension-specific communicative functions for the Task dimension. Dimension-specific functions are therefore defined only for dialogue control acts, which manage the interaction. Tables 8.6 and 8.7 list the pre-

Table 8.5: Formalised preconditions for the action discussion communicative functions defined in DIT (S = sender; A = addressee; α = action; σ = certainty; C_α = condition).

Communicative function	Preconditions
Instruct	$Want(S, CommitDo(A, \alpha, C_\alpha))$ $Assume(S, CanDo(A, \alpha))$ $Assume(S, WillDo(A, \alpha, C_\alpha))$
Request	$Want(S, [WillDo(A, \alpha, C_\alpha) \rightarrow CommitDo(A, \alpha, C_\alpha)])$ $Assume(S, CanDo(A, \alpha))$
Offer	$Want(S, Bel(A, WillDo(S, \alpha, C_\alpha)))$ $WillDo(S, \alpha, C_\alpha)$
Promise	$CommitDo(S, \alpha, C_\alpha)$ $Want(S, Bel(A, CommitDo(S, \alpha, C_\alpha)))$ $Bel(S, Interest(A, \alpha))$
Suggest	$Want(S, ConsidDo(A, \alpha, A, C_\alpha))$ $Assume(S, CanDo(A, \alpha))$ $Bel(S, Interest(A, \alpha))$
Address Request	$Bel(S, Want(A, [WillDo(S, \alpha, C_\alpha) \rightarrow CommitDo(S, \alpha, C_\alpha)]))$ $ConsidDo(S, \alpha, S, C_\alpha)$ $Bel(S, Assume(A, CanDo(S, \alpha)))$
AcceptRequest	$Bel(S, Want(A, [WillDo(S, \alpha, C_\alpha) \rightarrow CommitDo(S, \alpha, C_\alpha)]))$ $CommitDo(S, \alpha, C_\alpha)$ $Bel(S, Assume(A, CanDo(S, \alpha)))$
DeclineRequest	$Bel(S, Want(A, [WillDo(S, \alpha, C_\alpha) \rightarrow CommitDo(S, \alpha, C_\alpha)]))$ $CommitRefrain(S, \alpha, C_\alpha)$ $Bel(S, Assume(A, CanDo(S, \alpha)))$
AddressOffer	$ConsidDo(S, \alpha, A, C_\alpha)$ $Bel(S, WillDo(A, \alpha, C_\alpha))$ $Bel(S, Want(A, Bel(S, WillDo(A, \alpha, C_\alpha))))$
AcceptOffer	$Want(S, CommitDo(A, \alpha, C_\alpha))$ $Bel(S, WillDo(A, \alpha, C_\alpha))$ $Bel(S, Want(A, Bel(S, WillDo(A, \alpha, C_\alpha))))$
DeclineOffer	$Want(S, CommitRefrain(A, \alpha, C_\alpha))$ $Bel(S, WillDo(A, \alpha, C_\alpha))$ $Bel(S, Want(A, Bel(S, WillDo(A, \alpha, C_\alpha))))$
AddressSuggest	$ConsidDo(S, \alpha, S, C_\alpha)$ $Bel(S, Want(A, ConsidDo(S, \alpha, A, C_\alpha)))$ $Bel(S, Interest(A, \alpha))$ $Bel(S, Assume(A, CanDo(S, \alpha)))$
AcceptSuggest	$CommitDo(S, \alpha, C_\alpha)$ $Bel(S, Want(A, ConsidDo(S, \alpha, A, C_\alpha)))$ $Bel(S, Interest(A, \alpha))$ $Bel(S, Assume(A, CanDo(S, \alpha)))$
DeclineSuggest	$CommitRefrain(S, \alpha, C_\alpha)$ $Bel(S, Want(A, ConsidDo(S, \alpha, A, C_\alpha)))$ $Bel(S, Interest(A, \alpha))$ $Bel(S, Assume(A, CanDo(S, \alpha)))$

conditions of the DIT dimension-specific Auto- and Allo-Feedback functions respectively.

Table 8.8 lists the preconditions of the DIT functions specific for Interaction Management

Table 8.6: Formalised preconditions for the Auto-Feedback communicative functions defined in DIT (S = sender; A = addressee; fs = functional segment; da = dialogue act).

Communicative function	Preconditions
positive feedback	$Bel(S, Pos.Processed(S, fs))$
(underspecified)	$Want(S, Bel(A, Pos.Processed(S, fs)))$
positive execution	$Bel(S, Executed(S, da))$
	$Want(S, Bel(A, Executed(S, da)))$
positive evaluation	$Bel(S, Evaluated(S, da))$
	$Want(S, Bel(A, Evaluated(S, da)))$
positive interpretation	$Bel(S, Interpreted(S, fs))$
	$Want(S, Bel(A, Interpreted(S, fs)))$
positive perception	$Bel(S, Perceived(S, fs))$
	$Want(S, Bel(A, Perceived(S, fs)))$
positive attention	$Bel(S, Attended(S, fs))$
	$Want(S, Bel(A, Attended(S, fs)))$
negative feedback	$Bel(S, Neg.Processed(S, fs))$
(underspecified)	$Want(S, Bel(A, Neg.Processed(S, fs)))$
negative execution	$Bel(S, \neg Executed(S, da))$
	$Want(S, Bel(A, \neg Executed(S, da)))$
negative evaluation	$Bel(S, \neg Evaluated(S, da))$
	$Want(S, Bel(A, \neg Evaluated(S, da)))$
negative interpretation	$Bel(S, \neg Interpreted(S, fs))$
	$Want(S, Bel(A, \neg Interpreted(S, fs)))$
negative perception	$Bel(S, \neg Perceived(S, fs))$
	$Want(S, Bel(A, \neg Perceived(S, fs)))$
negative attention	$Bel(S, \neg Attended(S, fs))$
	$Want(S, Bel(A, \neg Attended(S, fs)))$

Table 8.7: Formalised preconditions for the Allo-Feedback communicative functions defined in DIT (S = sender; A = addressee; fs = functional segment; da = dialogue act).

positive feedback	$Bel(S, Pos.Processed(A, fs))$
(underspecified)	$Want(S, Bel(A, POs.Processed(A, fs)))$
positive execution	$Bel(S, Executed(A, da))$
	$Want(S, Bel(A, Executed(A, da)))$
positive evaluation	$Bel(S, Evaluated(A, da))$
	$Want(S, Bel(A, Evaluated(A, da)))$
positive interpretation	$Bel(S, Interpreted(A, fs))$
	$Want(S, Bel(A, Interpreted(A, fs)))$
positive perception	$Bel(S, Perceived(A, fs))$
	$Want(S, Bel(A, Perceived(A, fs)))$
positive attention	$Bel(S, Attended(A, fs))$
	$Want(S, Bel(A, Attended(A, fs)))$
negative feedback	$Bel(S, Neg.Processed(A, fs))$
(underspecified)	$Want(S, Bel(A, Neg.Processed(A, fs)))$
negative execution	$Bel(S, \neg Executed(A, da))$
	$Want(S, Bel(A, \neg Executed(A, da)))$
negative evaluation	$Bel(S, \neg Evaluated(A, da))$
	$Want(S, Bel(A, \neg Evaluated(A, da)))$
negative interpretation	$Bel(S, \neg Interpreted(A, fs))$
	$Want(S, Bel(A, \neg Interpreted(A, fs)))$
negative perception	$Bel(S, \neg Perceived(A, fs))$
	$Want(S, Bel(A, \neg Perceived(A, fs)))$
negative attention	$Bel(S, \neg Attended(A, fs))$
	$Want(S, Bel(A, \neg Attended(A, fs)))$
feedback elicitation (underspecified)	$Want(S, KnowVal(S, Pos.Processed(A, da)))$
elicit execution	$Want(S, KnowVal(S, Executed(A, da)))$
elicit evaluation	$Want(S, KnowVal(S, Evaluated(A, da)))$
elicit interpretation	$Want(S, KnowVal(S, Interpreted(A, fs)))$
elicit perception	$Want(S, KnowVal(S, Perceived(A, fs)))$
elicit attention	$Want(S, KnowVal(S, Attended(A, fs)))$

dimensions.[4]

Effects of qualifiers

Three categories of communicative function qualifiers are defined: certainty, conditionality and sentiment (see Chapter 4, Section 4.4). The attachment of qualifiers to communicative function has certain consequences for the update effects of such functions. Communicative function qualifiers are semantically defined as making the information state updates of the communicative functions that they qualify more specific. For example, information-providing functions can be qualified as to how certain the speaker is of the correctness of the information that he provides. The 'uncertain' qualifier has the effect that the addressee's information is updated with the information that the speaker has a weak belief (as opposed to a firm belief) that the information he provides is correct. An unqualified Inform act, for example, has the preconditions $Bel(S, p, \sigma)$ and $Want(S, Bel(A, p, \sigma))$, expressing that S holds the belief that p with certainty σ, and has the goal that the same will be the case for A. The qualifier 'uncertain' specifies σ to have the value 'weak': $Bel(S, p, weak)$; $Want(S, Bel(A, p, weak))$, and the qualifier 'certain' the value 'firm': $Bel(S, p, firm)$; $Want(S, Bel(A, p, firm))$.

Action-discussion functions may be qualified with respect to conditionality. When an agent A is considering to do an action α under certain condition(-s) C_α, this is represented by $ConsidDo(A, \alpha, C_a)$ where C_a may be empty, i.e. the universally true condition '$\top$'. To represent situations where the speaker considers an action to be performed by another participant (e.g. *Maybe better to remove this panel if you can*) the four-place predicate $ConsidDo(X, \alpha, Y, C_a)$ is defined.

Sentiment qualifiers add information about the speaker's attitudinal and emotional state, such as pleasure, surprise, annoyance or irritation.

Formally, qualifiers come in two varieties, called *'q-specifiers'* and *'q-additives'*, that have a different semantic effect (Bunt, 2011). Q-specifiers make the preconditions of the communicative function that they qualify more specific. Certainty and conditionality qualifiers are both q-specifiers.[5] Q-additives enrich a communicative function with additional information. Sentiment qualifiers are q-additives.

For the semantics of qualified communicative functions we thus have three possible cases to consider, where f_i is an unqualified communicative function: (a) $\langle f_i; qs_j \rangle$ where qs_j is a q-specifier; (b) $\langle f_i; qa_k \rangle$ where qa_k is a q-additive; and (c) $\langle f_i; qs_j; qa_k \rangle$ where qs_j is a q-specifier and qa_k is a q-additive.

We consider two examples. The first illustrates the semantics of an answer, qualified as uncertain, as in (81):

(81) A: When is the next meeting?
B: I think on Friday December 3.

Applied to participants B and A in (81), the understanding of the answer has the effect that A's pending semantic context is extended with the following pieces of information (where M_d abbreviates the description of the date of the next meeting, and $M - Friday - 12.3$ the proposition that the next meeting will be on Friday, December 3):

[4] Since Social Obligation Management acts do not play an important role in this thesis, we do not attempt to formalise them.

[5] In Petukhova and Bunt (2010b) a third q-specifier was distinguished, concerned with partiality, but we now believe that partiality is better treated in a different way, see also Chapter 4.

Table 8.8: Formalised preconditions for the Interaction Management communicative functions defined in DIT. (S = sender; A = addressee; fs = functional segment; fs_p = part of a functional segment; t = topic).

Communicative function	Preconditions
Turn Take	$Bel(S, \neg Next\_Speaker(S) \wedge \neg Next\_Speaker(A))$ $Want(S, Next\_Speaker(S))$
Turn Accept	$Bel(S, Want(A, Next\_Speaker(S)))$ $Want(S, Next\_Speaker(S))$
Turn Grab	$Bel(S, Current\_Speaker(A))$ $Bel(S, Want(A, Next\_Speaker(A)))$ $Want(S, Next\_Speaker(S))$
Turn Keep	$Bel(S, Current\_Speaker(S))$ $Want(S, Next\_Speaker(S))$
Turn Assign	$Bel(S, Current\_Speaker(S))$ $Want(S, Next\_Speaker(A))$
Turn Release	$Bel(S, Current\_Speaker(S))$ $Want(S, \neg Next\_Speaker(S))$
Stalling	$Bel(S, Time\_Need(S, small))$ $Want(S, Bel(A, Time\_Need(S, small)))$
Pausing	$Bel(S, Time\_Need(S, substantial)))$ $Want(S, Bel(A, Time\_Need(S, substantial)))$
ContactCheck	$Want(S, Ready(A))$
ContactIndication	$Want(S, KnowVal(A, Ready(S)))$
Opening	$CommitDo(S, Open-dialogue, \top)$
Pre-closing	$CommitDo(S, Close-dialogue, \top)$
Topic introduction	$Want(S, Next\_Topic(t1))$
Topic shift	$Bel(S, Current\_Topic(t1))$ $Want(S, Next\_Topic(t2))$
Error signal	$Bel(S, Mistake(S, fs_p))$ $Want(S, Bel(A, Mistake(S, fs_p)))$
Retract	$Bel(S, Mistake(S, fs_p))$ $Want(S, Delete(S, fs_p))$
Self Correction	$Bel(S, Mistake(S, fs_p))$ $Want(S, Substitute(S, (fs_p, fs_c)))$
Completion	$Want(S, Concatenate(A, (fs_p, fs_c)))$
Correct Misspeaking	$Bel(S, Mistake(A, fs_p))$ $Want(S, Substitute(A, (fs_p, fs_c)))$

(82) 1. $Bel(A, Bel(B, \textit{M-Friday-12.3}, \textit{weak}))$, or equivalently: $WBel(B, \textit{M-Friday-12.3})$; i.e., A believes that B holds the uncertain belief that the next meeting is on Friday-12.3;
2. $Bel(A, Want(B, WBel(A, \textit{M-Friday-12.3})))$, i.e. A believes that B has the goal that A also holds this uncertain belief;
3. $Bel(A, Assume(B, Want(A, KnowVal(A, M_d))))$, i.e. A believes that B assumes that A wants to know when is the next meeting.
4. $Bel(A, Assume(B, Assume(A, KnowVal(B, M_d))))$: A believes that B assumes that A assumes that B knows when is the next meeting.

Second, example (83) illustrates the semantics of an conditional Accept Offer with a happy sentiment:

(83) A: Would you like to have some coffee?

B: If you have, that would be wonderful!

Applied to the participants A and B in (83), the action $coffee$, and the condition *A-able-arrange-coffee*, understanding this Accept Offer has the effect that A's pending semantic context is extended with the following pieces of information:

(84) 1. $Bel(A, Want(B, CommitDo(A, coffee, \textit{A-able-arrange-coffee}))$, i.e., A holds the belief that B wants A to arrange coffee for B if A is able to do so.
2. $Bel(A, Bel(B, WillDo(A, coffee, \textit{A-able-arrange-coffee}))$, i.e. A believes that B believes that A has the goal to arrange coffee for B if B wants that.
3. $Bel(A, Bel(B, Want(A, Bel(B, WillDo(A, coffee, \textit{A-able-arrange-coffee})))))$, i.e. A believes that B believes that A wants B to believe that A is willing to arrange coffee for B.
4. $Bel(A, HAPPY(B, coffee))$: A believes that B is happy to have some coffee.

In other words, the Task component of A's pending context is extended with the beliefs representing B's wish that A commits himself to arrange coffee if he is able to do so; that A is willing to do so; and that A wants B to believe that. Moreover, the fact that B is happy to get coffee is represented in the cognitive component of A's pending context.

8.3 Context-driven dialogue act generation

The information state of an addressee of a dialogue act changes when he understands the speaker's behaviour. When an addressee merely pays attention he believes that his communicative partner has a speaker role, and has said or done something. These beliefs are added to his Linguistic Context. When an addressee perceived what was said or performed, then his context is updated with this information, e.g. beliefs about the verbatim form of an utterance or about visible gestures are recorded in his Linguistic Context. When he understands the speaker's behaviour this means that he is able to identify functional segments and their intended interpretation as dialogue acts. Understanding that a certain dialogue act is performed means believing that the preconditions which are characteristic for that dialogue act hold. These beliefs are thus added to an addressee's pending context. This is a *buffering* stage. If one or more buffered beliefs are inconsistent with beliefs already present in the context, then these beliefs cannot simply be added to that context. The beliefs that correspond to understanding that a

certain dialogue act is performed, are therefore always first buffered and evaluated for consistency with beliefs already present. Beliefs at this stage may be *accepted*, and weak beliefs may be *strengthened*. If no inconsistencies are detected, the new beliefs can be added to the main context. Finally, an addressee tries to execute activities that correspond to *adopting* beliefs conveyed in the speaker's acts. Thus, in the processing of the incoming utterance several stages can be distinguished: awareness, recording, buffering, acceptance, and adoption.

Each processing stage corresponds to the application of a number of mechanisms for updating the addressee's context. Following Bunt (2005), Morante (2007) defines the following general context update mechanisms:

- *Creation*: an interlocutor introduces a belief as the effect of assigning an interpretation to what has been said by another interlocutor. Creation has two stages: (1) addition of precondition to the pending context; and (2) acceptance of beliefs and addition of accepted elements to the main context;
- *Adoption*: an interlocutor incorporates beliefs of an other interlocutor as beliefs of his own;
- *Cancellation*: a belief or goal is cancelled because it does not apply any more, or a goal has been achieved or has been understood to be unachievable;
- *Strengthening*: an expectation, or 'weak belief' becomes a firm belief because sufficient supporting evidence for the belief becomes available.

Additionally to these general mechanisms, that are applicable to the processing of any dialogue act, we need more refined specific mechanisms, namely feedback mechanisms and update mechanisms concerning turn allocation. We define them as follows:

- *Identification*: an interlocutor generates the goal to report on his success of processing a contribution;
- *Turn Allocation*: if an interlocutor has a goal which could be achieved with the help of a certain dialogue act, then he generates the goal to occupy the speaker role, in order to be able to perform that dialogue act.

Context update mechanisms provide the specification of how the information states of dialogue participants change during the dialogue as the result of mutual understanding. A speaker's expectation of being understood and believed is modelled in DIT by the speaker having a 'weak belief' that the addressee believes that the preconditions hold and the content of the dialogue act is true. Thus, additional communicative effects occur as follows (see Morante, 2007):

- *Understanding effects*: the effects of an addressee understanding a dialogue utterance of the speaker, its implied feedback effects on the previous utterance of the addressee and other implicated effects. The addressee upon understanding the utterance believes that the corresponding preconditions apply for the speaker;
- *Expected understanding effects*: the effects of the speaker expecting (in terms of weak beliefs) that the understanding effects for the addressee take place. The assumption by both interlocutors of expected understanding leads to the mutual belief that the speaker weakly believes that the understanding effects occur;

- *Adoption effects*: the effects of the addressee incorporating the information presented by the speaker in his own information state;
- *Expected adoption effects*: effects of the speaker expecting the adoption effects for the addressee to take place. The assumption by both interlocutors of expected adoption leads to the mutual belief that the speaker weakly believes that the addressee adopts the information presented.

Update mechanisms and communicative effects specify how an addressee's context is changed, and how the new context is created for the addressee to respond to a dialogue act. Take for example a Propositional Question as in the following dialogue example:[6]

(85) Ian: Is this a large sample?
Marc: Well, this is not large sample

The preconditions for the performance of a Propositional Question in a utterance u of the speaker U addressed to S (with Ian as U and Mark as S), and with semantic content p (= 'this is a large sample') are as given in Table 8.9: $u01 = \mathit{Want}(U, \mathit{KnowVal}(U, p)); u02 = \mathit{Assume}(U, \mathit{KnowVal}(S, p))$. The goal that the speaker is trying to achieve is $\mathit{KnowVal}(U, p)$. If the addressee S understands the utterance as a Propositional Question and acts cooperatively and rationally, then these conditions give rise to giving an answer to this question, unless processing problems occur like in example (86) below. In the case of successful understanding the addressee's context is updated as follows: S believes that U wants to know the truth value of the proposition p and S believes that U assumes that S knows the truth value of p. These beliefs are represented in S's context as *s6a* and *s6b*. Again acting cooperatively S adopts U's goal and therefore $\mathit{Want}(S, \mathit{KnowVal}(U, p))$. Being a rational agent, S provides an answer containing S's knowledge that $\neg p$.

Additionally, the expectation of the speaker of the Propositional Question that the addressee perceives and understands his behaviour, corresponding to the functional segment fs_1, gives rise to updates in the Cognitive Contexts of both communicative partners, and if the addressee S has no problems processing U's behaviour the goal is generated to report this. A positive auto-feedback act is planned as a candidate act to be generated. Since the answer, that is also is generated as a candidate dialogue act, entails positive understanding, it is not necessary to express the positive auto-feedback act explicitly. This might suggest that giving positive feedback is not necessary when it is entailed or implicated by other dialogue acts, as is the case of all responsive acts (such as Answer, Accept or Decline Request, Offer or Suggestion). Entailed positive feedback can be provided explicitly for strategic reasons, e.g. when automatic speech recognition is used to avoid misunderstandings. The issues why and when to generate explicit positive feedback will be discussed in some detail in Section 8.4. In our example (85) the speaker S decides to report the entailed positive auto-feedback explicitly by repeating part of U's question.

[6] From the AMI pilot meeting corpus - pilot meeting 2.

Table 8.9: Example of updated context for a pair Propositional Question - Answer. (LC = Linguistic Context; SC = Semantic Context; CC = Cognitive Context; prec = preconditions; impl = by implication; du = dialogue utterance; da = dialogue act; fs = functional segment; D = dimension; CF = communicative function; exp.und = expected understanding; und = understanding; exp.ad = expected adoption; ad = adoption; MBel = mutually believed)

Context	num	source/ *role*	S's context	num	source/ *role*	U's context
SC				u01 u02	*prec*	$Want(U, KnowVal(U,p))$ $Assume(U, KnowVal(S,p))$
	s1		$Bel(S, Current\_Speaker(U))$	u1		$Bel(U, Current\_Speaker(U))$
	$fs_1 : du1$	*latest*	$\langle t_1 = is, t_2 = this, t_3 = a, t_4 = large, t_5 = sample \rangle$	$fs_1 : du1$	*latest*	$\langle t_1 = is, t_2 = this, t_3 = a, t_4 = large, t_5 = sample \rangle$
	$fs_1 : da_1$	*D;CF*	Task; PropositionalQuestion Speaker:U; Addressee:S	$fs_1 : da_1$	*D;CF*	Task; PropositionalQuestion Speaker:U; Addressee:S
		sem_content	$p = S1$ *is a large sample*		*sem_content*	$p = S1$ *is a large sample*
	$fs_1 : da_2$	*default*	TurnM.;TurnAssign Speaker:U; Addressee:S	$fs_1 : da_2$	*default*	TurnM.;TurnAssign Speaker:U; Addressee:S
	s2	exp.und: $fs_1 : da_2$	$Bel(S, MBel(\{S,U\}, WBel(U, Want(U, Next\_Speaker(S))))$	u2	exp.und: $fs_1 : da_2$	$Bel(S, MBel(\{S,U\}, WBel(U, Next\_Speaker(S))))$
	s3	und:u2	$Bel(S, Want(U, Next\_Speaker(S))$			
	s4	ad:da_2	$Want(S, Next\_Speaker(S))$			
	s00	s4	$Want(S, Bel(U, Next\_Speaker(S)))$			
SC	s5a	exp.und: $fs_1 : da_1$	$Bel(S, MBel(\{S,U\}, WBel(U, Bel(S, Want(U, KnowVal(U,p)))$	u3a	exp.und: $fs_1 : da_1$	$Bel(U, MBel(\{S,U\}, WBel(U, Bel(S, Want(U, KnowVal(U,p)))$
	s5b	exp.und: $fs_1 : da_1$	$Bel(S, MBel(\{S,U\}, WBel(U, Bel(S, Assume(U, KnowVal(S,p)))$	u3b	exp.und: $fs_1 : da_1$	$Bel(U, MBel(\{S,U\}, WBel(U, Bel(S, Assume(U, KnowVal(S,p)))$
	s6a	und:u3a	$Bel(S, Want(U, KnowVal(U,p)))$			
	s6b	und:u3b	$Bel(S, Assume(U, KnowVal(S,p)))$			
	s7	ad:da_1	$Want(S, KnowVal(U,p))$			
	s01	s7	$Bel(S, \neg p))$			
CC	s8	exp.und: $fs_1 : du1$	$Bel(S, MBel(\{S,U\}, WBel(U, Interpreted(S, du1)))$	u4	exp.und: $fs_1 : du1$	$Bel(U, MBel(\{S,U\}, WBel(U, Interpreted(S, du1)))$
	s9	und:u4	$Bel(S, Interpreted(S, du1))$			
	s02	s9	$Want(S, Bel(U, Interpreted(S, du1)$			

Context	num	source/ *role*	S's context	num	source/ *role*	U's context
LC	da_3	*plan:s00*	Turn Accept Speaker:S; Addressee:U antecedent: da_2			
	da_4	*plan:s02*	Auto-F.;Interpretation Speaker:S; Addressee:U antecedent: fs_1			
	da_5	*plan:s01*	Task; Answer Speaker:S; Addressee:U antecedent: da_1			
LC	s10		$Bel(S, Current\_Speaker(S))$	u5		$Bel(U, Current\_Speaker(S))$
	$fs_2 : du2$	*latest*	$\langle t_1 = well \rangle$	$fs_2 : du2$	*latest*	$\langle t_1 = well \rangle$
	$fs_2 : da_3$	*D;CF*	TurnM.;TurnAccept Speaker:S; Addressee:U antecedent: $fs_1 : da_2$	$fs_2 : da_3$	*D;CF*	TurnM.;TurnAccept Speaker:S; Addressee:U antecedent: $fs_1 : da_2$
	$fs_3 : du3$	*latest*	$\langle t_2 = this, t_3 = is, t_5 = large, t_6 = sample \rangle$	$fs_3 : du3$	*latest*	$\langle t_2 = this, t_3 = is, t_5 = large, t_6 = sample \rangle$
	$fs_3 : da_4$	*D;CF*	Auto-F.;Pos.Interpretation Speaker:S; Addressee:U antecedent: $fs_1 : du1$	$fs_3 : da_4$	*D;CF*	Auto-F.;Pos.Interpretation Speaker:S; Addressee:U antecedent: $fs_1 : du1$
	$fs_4 : du4$	*latest*	$\langle t_2 = this, t_3 = is, 7_4 = not,$ $t_5 = large, t_6 = sample \rangle$	$fs_4 : du4$	*latest*	$\langle t_2 = this, t_3 = is, 7_4 = not,$ $t_5 = large, t_6 = sample \rangle$
	$fs_4 : da_5$	*D;CF*	Task; Answer Speaker:S; Addressee:U	$fs_4 : da_5$	*D;CF*	Task; Answer Speaker:S; Addressee:U
		sem_content	$\neg p$ antecedent: $fs_1 : da_1$		*sem_content*	$\neg p$ antecedent: $fs_1 : da_1$

Participants in dialogue do not always process speaker utterances successfully. Consider the situation when the same propositional question from example (85) is not perceived quite successfully:

(86) Ian: Is this a large sample?
Marc: Is this a what?

Table 8.10 shows what effects are created and how plans to perform certain acts are constructed, when the speaker (with Marc for S) encounters processing problems at the level of perception.

As in the case of example (85), the speaker U expects S to understand his utterance as a propositional question. S understands that a propositional question is performed, but fails to understand what the question is about. This creates the goal for S to perform an Auto-Feedback Set Question to find out what it was that he did not perceive successfully. The processing problem that S is experiencing is serious enough to prevent further task performance and should be resolved first; this problem is reported accordingly.

In examples (85) and (86), the question act has an accompanying default Turn Assigning function: $Want(U, Next\_Speaker(S))$. The understanding of this turn management act affects the addressee's Linguistic Context, and in order to provide an answer S generates the goal to take the turn (turn accepting is seen as a side-effect of providing an answer). In example (85) he explicitly indicates taking the turn by using the discourse marker *'well'*.

When the addressee understands that a turn assign act is performed, this puts pressure on him to accept the speaker role. When no turn final act is performed and the performed act has no implicated turn final function (e.g. Inform, Answer, Suggest and many others), participants operate on the basis of their beliefs about the availability of the speaker role, and of how eager they are to have the speaker role. Every dialogue act in spoken form requires the speaker to have the speaker role. Hence he should be able to take or accept that role or else he will have to set up the goal to obtain the speaker role, possibly leading to a Turn Grab act. Like in the case of explicit positive auto-feedback acts the generation of explicit turn management acts heavily depends on the dialogue setting and may lead to different dialogue strategies (see Section 8.4).

To sum up, the defined update mechanisms specify how an addressee's context is changed and new contexts are created. Communicative effects as described above can be used to decide what update mechanisms should be applied. Expected understanding and adoption effects together with the agent's goals, motivated by underlying task and general principles of cooperativity and rationality, give rise to the generation of dialogue acts in multiple dimensions. Generation of dialogue contributions not only involves decisions about which dialogue act(-s) are motivated by the preceding and current context, but also making choices among licensed dialogue acts and deciding how to express combinations of chosen dialogue acts verbally and non-verbally. The next section describes procedures for the selection of candidate dialogue acts and considerations how the selected dialogue acts can be used to generate multifunctional utterances.

Table 8.10: Example of updated context for a pair Propositional Question-Propositional Answer. (LC = Linguistic Context; SC = Semantic Context; CC = Cognitive Context; prec = preconditions; impl = by implication; du = dialogue utterance; da = dialogue act; fs = functional segment; D = dimension; CF = communicative function; exp.und = expected understanding; und = understanding; exp.ad = expected adoption; ad = adoption; MBel = mutually believed)

Context	num	source *role*	S's context	num	source *role*	U's context
SC				u01	*prec*	$Want(U, KnowVal(U,p))$
				u02		$Assume(U, KnowVal(S,p))$
LC				u03	*prec*	$Bel(U, Next\_Speaker(U))$
LC	s1		$Bel(S, Current\_Speaker(U))$	u1		$Bel(U, Current\_Speaker(U))$
	$fs_1 : du1$	*latest*	$\langle t_1 = is, t_2 = this, t_3 = a,$ $t_4 = x, t_5 = x\rangle$	$fs_1 : du1$	*latest*	$\langle t_1 = is, t_2 = this, t_3 = a,$ $t_4 = large, t_5 = sample\rangle$
	$fs_1 : da_1$	*D;CF*	Task; PropositionalQuestion Speaker:U; Addressee:S	$fs_1 : da_1$	*D;CF*	Task; PropositionalQuestion Speaker:U; Addressee:S
		sem_content	x		*sem_content*	$p = S1$ *is a large sample*
	$fs_1 : da_2$	default	Turn-M.;TurnAssign Speaker:U; Addressee:S	$fs_1 : da_2$	default	Turn-M.;TurnAssign Speaker:U; Addressee:S
	s2	exp.und: $fs_1 : da_2$	$Bel(S, MBel(\{S,U\}, WBel(U,$ $Next\_Speaker(S))))$	u2	exp.und: $fs_1 : da_2$	$Bel(S, MBel(\{S,U\}, WBel(U,$ $Next\_Speaker(S))))$
	s3	und:u2	$Bel(S, Want(U, Next\_Speaker(S)))$			
	s4	ad:da_2	$Bel(S, Want(S, Next\_Speaker(S)))$			
	s00	s4	$Want(S, Bel(U, Next\_Speaker(S)))$			
SC	52a	exp.und: $fs_1 : da_1$	$Bel(S, MBel(\{S,U\}, WBel(U,$ $Bel(S, \exists x.Want(U, KnowVal(U,x)))$	u3a	exp.und: $fs_1 : da_1$	$Bel(U, MBel(\{S,U\}, WBel(U,$ $Bel(S, \exists x.Want(U, KnowVal(U,x)))$
	s5b	exp.und: $fs_1 : da_1$	$Bel(S, MBel(\{S,U\}, WBel(U,$ $Bel(S, \exists x.Assume(U, KnowVal(S,x)))$	u3b	exp.und: $fs_1 : da_1$	$Bel(U, MBel(\{S,U\}, WBel(U,$ $Bel(S, \exists x.Assume(U, KnowVal(S,x)))$
	s6a	und:u3a	$Bel(S, \exists x.Want(U, KnowVal(U,x)))$			
	s6b	und:u3b	$Bel(S, \exists x.Assume(U, KnowVal(S,x)))$			
	s7	ad:da_1	$\iota x.Want(S, KnowVal(U,x))$			
	s01	s7	$Want(S, KnowVal(S, \iota x.Want(U,$ $KnowVal(U,x))))$			
CC	s8	exp.und: $fs_1 : du1$	$Bel(S, MBel(\{S,U\}, WBel(U,$ $Interpreted(S, du1)))$	u4	exp.und: $fs_1 : du1$	$Bel(U, MBel(\{S,U\}), WBel(U,$ $Interpreted(S, du1)))$
	s9a	und:u4	$Bel(S, \neg Perceived(S, (t_4, t_5)))$			
	s9b	und:u4	$Bel(S, Perceived(S, (t_1 - t_4)))$			
	s02	s9a	$Want(S, Bel(U, \neg Perceived(S, (t_4, t_5))))$			

LC	da_3	*plan:s00*	TurnM.;Turn Accept Speaker:S; Addressee:U antecedent: da_2			
	da_4	*plan:s02*	Auto-F.;Pos.Perception Speaker:S; Addressee:U antecedent: $fs_1 : du1 : t_4, t_5$			
	da_5	*plan:s02*	Auto-F.;Neg.Perception Speaker:S; Addressee:U antecedent: $fs_1 : du : t_1 - t_3$			
	da_6	*plan:s01*	Auto-F.; SetQuestion Speaker:S; Addressee:U antecedent: $fs_1 : du : t_1 - t_3$			
LC	s10		$Bel(S, Current\_Speaker(S))$	u5		$Bel(U, Current\_Speaker(S))$
	$fs_2 : du2$	*latest*	$\langle t_9 = is, t_1 0 = this, t_1 1 = a, t_1 2 = what \rangle$	$fs_2 : du2$	*latest*	$\langle t_9 = is, t_1 0 = this, t_1 1 = a, t_1 2 = what \rangle$
	$fs_2 : da_4$	*D;CF*	Auto-F.;Pos.Perception Speaker:S; Addressee:U antecedent: $fs_1 : du1 : t_1 - t_3$	$fs_3 : da_4$	*D;CF*	Auto-F.;Pos.Perception Speaker:S; Addressee:U antecedent: $fs_1 : du1 : t_1 - t_3$
	$fs_2 : da_6$	*D;CF*	Auto-F.;SetQuestion Speaker:S; Addressee:U antecedent: $fs_1 : du1 : t_4, t_5$	$fs_2 : da_5$	*D;CF*	Auto-F.;SetQuestion Speaker:S; Addressee:U antecedent: $fs_1 : du1 : t_4, t_5$
	$fs_2 : da_3$	*side-effect*	TurnM.;TurnAccept Speaker:S; Addressee:U antecedent: $fs_1 : da_2$	$fs_2 : da_3$	*side-effect*	TurnM.;TurnAccept Speaker:S; Addressee:U antecedent: $fs_1 : da_2$

8.4 Selection of dialogue acts for generation

Generated dialogue act candidates for continuing the dialogue are stored in the dialogue future part of the Linguistic Context. Dialogue acts pertaining to different dimensions can be generated independently, but for their order of performance and their combination, the relative importance of the dimensions at the given point in the dialogue has to be taken into account. Selecting dialogue acts for realisation in a dialogue utterance involves logical, pragmatic, strategic and linguistic considerations.

Candidate dialogue acts may be in logical or pragmatic conflict with each other. The entailment relations that may exist between dialogue acts give rise to logical constraints, whereas implicatures between acts give rise to pragmatic constraints. Both types of constraints should be taken into account to avoid inconsistencies in dialogue act combinations, ensuring rational behaviour. Section 8.4.1 discusses such constraints.

Some of the dialogue act candidates may have high priority and should be generated at once, such as those that signal a serious communicative problem; others may be postponed Some will already be performed by implication through the performance of other candidate acts. The assignment of priorities to dialogue act candidates will be discussed in Section 8.4.2.

For a set of dialogue acts that have no conflicts arising from entailments or implicatures, there may be constraints that depend on the particular setting in which the dialogue system is used. Such constraints offer a way to implement different dialogue strategies and interactive styles. Section 8.4.3 will illustrate several cases where the generation of multiple dialogue act candidates offers such options.

Finally, combinations of dialogue acts should be selected that can actually be realised in multifunctional utterances. Some combinations of dialogue acts may be hard or strange to express in a single utterance in natural language under consideration. Section 8.4.4 discusses linguistic constraints on the expressibility of dialogue act combinations in a single functional segment, in an utterance, or in a sequence of segments or utterances within a turn unit, and indicates some possibilities for the use of multiple modalities.

8.4.1 Constraints on the combinations of dialogue acts

The DIT^{++} tag set has 26 general-purpose functions, which may be used in every dimension, and 56 dimension-specific functions. The distribution of function across dimensions is as shown in Figure 8.3. A functional segment may have a communicative function in each of the 10 dimensions, or only in 9 of them, or only in 8 of them, or ... in only one. Hence the total number of possible combinations is the sum of the possible combinations of 10 tags, of 9 tags, of 8 tags, ... of single tags. The number of possible combinations of 10 tags is $26 \times 38 \times 44 \times 32 \times 28 \times 28 \times 32 \times 29 \times 28 \times 36 = 1.02 \times 10^{15}$; adding the number of possible combinations of nine tags or less gives a total of 1.04×10^{15}. In practice, it has been shown that 2 functions per segment is a realistic number when we only count functions expressed by virtue of utterance features and implicated functions (see Bunt, 2010). This gives us $(D_1 \times D_2 + D_1 \times D_3 + D_1 \times D_4 + ...) = 63,171$ possible dialogue act combinations.

We analysed these function combinations and determined whether there are additional constraints on their combination and whether these have a logical or a pragmatic origin. For each dialogue act the logical entailments were calculated and all dialogue act pairs were inspected for logical conflicts. Calculating the entailment relations among dialogue acts, which are defined through their preconditions (see Section 8.2) ensures completeness in the sense of finding

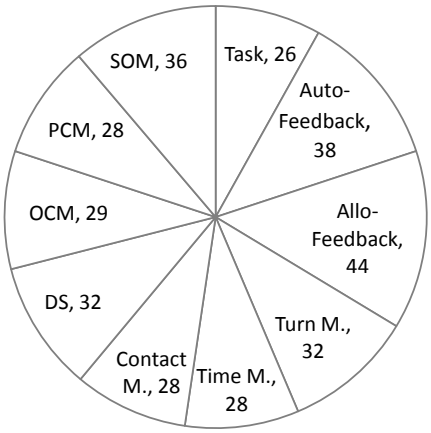


Figure 8.3: Distribution of functions across DIT dimensions.

all entailments between dialogue acts. While entailments depend solely on the definitions of communicative functions in terms of their preconditions, implicatures are pragmatic relations between a dialogue act and a condition that may be a precondition of another dialogue act, as will be illustrated below, and are a matter of empirical observation.

Logical constraints

From a logical point of view, two communicative functions cannot be applied to one and the same semantic content if they have logical conflicts in their preconditions or/and entailments. We analysed functional consistency pairwise between (1) preconditions of F_1 and F_2; (2) entailments of F_1 and F_2; (3) entailments of F_1 and preconditions of F_2 and vice versa.

The use of two functions (F_1 and F_2) applied to the same semantic content p is logically inconsistent if there is a proposition q which can be derived from the set of preconditions P_1 of F_1, while $\neg q$ can be derived from the preconditions P_2 of F_2. This is the case when we deal with alternative end-nodes in the tag set hierarchy. For example, one cannot accept and reject the same offer in one functional segment: Accept Offer requires that $Bel(S, WillDo(A, a))$; $Bel(S, CanDo(A, a))$; $Bel(S, Want(A, Bel(S, WillDo(A, a))))$ and $Want(S, CommitDo(A, p))$; for Reject Offer the same preconditions hold except for the last one which is $Want(S, CommitRefrain(A, a))$. Since $CommitRefrain(A, \alpha)$ entails $\neg CommitDo(A, \alpha)$, these sets of preconditions are inconsistent.

Two acts are also in conflict if the entailments of one are in logical conflict with the preconditions of the other. An obvious case is that of responsive dialogue acts and negative autofeedback. For example, in order to provide a correction the speaker needs to have paid attention, perceived and understood the relevant previous utterance.

Note that the combination of two conflicting acts is possible if they refer to different segments or acts in the previous discourse, i.e. if they have different *functional* or *feedback de-*

pendency relations, see Bunt (2010). For example:[7]

(87) U1: How long from Bath to Corning?
U2: An hour I think
S1: Two hours

Response S1 is an Answer to the Set Question in (U1) and a Correction of the Inform in (U2). The combination of Answer and Correction is logically not possible unless they have different relational antecedents, as in the case here.

Pragmatic constraints

Pragmatically speaking, two dialogue acts A_1 and A_2 are inconsistent in the following cases:

(88) (1) an implicated condition of A_1 blocks the performance of A_2;
(2) an implicated condition of A_1 is in conflict with an implicated condition of A_2.

Two dialogue acts cannot be combined in one segment if an implicature of one act makes the performance of another act impossible. For example, positive auto-feedback acts at the level of perception and lower do not satisfy the conditions for the speaker to be able to assist the addressee by providing a completion or a correction of the addressee's mistakes, because for being able to offer a completion or a correction it is not sufficient to pay attention and hear what was said, but understanding and evaluation are required, and signalling positive perception may implicate negative feedback at these higher processing levels.

Similarly, questions and requests have the default function that the speaker wants the addressee to have the next turn, hence the speaker does not want to have the next turn himself: $Want(S, Next\_Speaker(A))$, whereas such acts as Stalling or Pausing, but also acts like Self-Correction, Error Signalling and Retraction, implicate that the speaker wants to keep the turn: $Want(S, Next\_Speaker(S)$

As noted in (88.2), two acts cannot be combined in one segment if implicatures of one are in conflict with implicatures of another. For instance, Contact Check carries an implicature of negative perception of the partner's linguistic or nonverbal behaviour, whereas Opening carries an implicature of positive perception. Similarly, Partner Communication Management (PCM) acts are pragmatically inconsistent with dialogue acts like Opening, Self-Introduction, Greeting or Contact Check.

A general strategy for applying pragmatic constraints may be the following: if $A1$ is a dialogue act candidate, and $A1$ has an implicature $A2$, consider if $A2$ would be an admissible dialogue act candidate as well. If so, generate $A1$. If not, consider a dialogue act $A3$ which cancels the implicature $A2$ as a candidate dialogue act. For example, $A1$ = Thanking, $A2$ = PreClosing (phenomenon in two-party information seeking dialogues, see Bunt, 1992). If the speaker wants to thank the addressee, but not to close the dialogue, he should perform an act $A3$ to make clear that he wants to continue the dialogue.

Constraints for segment sequences

For sequential multifunctionality within turns there are fewer and softer constraints than for simultaneous multifunctionality. The combination of two mutually exclusive acts in a sequence is in principle possible, since a speaker can perform a dialogue act by mistake and subsequently

[7] From the TRAINS dialogue corpus.

correct himself, or can change his mind. Hence we may expect sequences of the following kind:

(89) 1. dialogue act A_1
2. retraction of A_1
3. dialogue act A_2

where A_1 and A_2 are conflicting. If logical or pragmatic conflicts are detected between candidate dialogue acts, they should not be generated in one segment, but marked as alternatives, that may be realised in separate segments possibly with segments in between that cancel or substitute the previous one.

Keizer et al. (2011) noticed that when two candidate dialogue acts conflict with each other, they can be combined in a sequence of functional segments with a discourse marker signalling the conflict, as in S3 example (90)[8].

(90) U1 : I see the send button
S1 : Okay
U2 : Where is the send button?
S2 : But you just told me you saw the send button!
S3 : The send button is on the bottom right, but you just told me you saw it!

After processing U2, the system detects a conflict between the user knowing where the send button is (from U1) and wanting to know where it is (from U2). This results in the generation of a candidate negative auto-feedback act, and at the same time an answer to the question in U2. In generating utterance S2 only the feedback act was selected, whereas alternatively both acts could be selected, resulting in S3. Which response is the best is a strategic matter, and depends on global dialogue conditions as well as on local conditions such as confidence scores propagated from an understanding module.

8.4.2 Assigning priorities to dialogue act candidates

Given a list of candidate dialogue acts addressing several dimensions we have choices (1) which acts to generate; and (2) in what order. These choices are not determined by dialogue act preconditions, but rather by theoretical and practical considerations. Priority constraints are potentially very complex and depend on numerous factors. Rather than trying to formulate priority rules, we focus on what to take into account when designing such rules. Based on the insights that we, obtained we will discuss some important points that matter when designing priority and strategic preference rules.

The priorities among independent dialogue acts in different dimensions depend on the type of communicative situation. For task-oriented dialogues, dialogue acts that address the task obviously have high priority. For example:[9]

(91) U1: How many boxcars of oranges are now in Bath?
S1: That'll be five
S1a: Uhm, okay, that'll be five
S1b: Uhm, okay, there are five boxcars of oranges in Bath

[8] From the DIAMOND dialogue corpus translated from Dutch.
[9] From the TRAINS dialogue corpus.

Utterance S1 is an answer to the question U1 about the dialogue task (planning optimal train transportation). Other dialogue acts that might be generated, include a turn acceptance act (S1a), and positive auto-feedback that the question is processed successfully (S1b), but it would be strange to not generate the task-related act.

In order to assess the relative importance of dialogue acts addressing different dimensions, empirical studies on MapTask data have been conducted and reported in (Włodarczak et al., 2010). Relationships were investigated between the occurrence of entailments among communicative functions, and dominance judgments were collected in an experiment in which participants rank utterance function in terms of their importance, assigning a numerical value to each function starting from '1' for the most important function. A strong tendency was found for entailed functions to be ranked lower than the entailing functions. This suggests that if $A1$ is a candidate dialogue act, and $A2$ is an entailment of $A1$, only A1 should be considered for generation. However, for positive auto-feedback the possibility of explicit generation of $A2$ may be considered, by means of repetitions, paraphrasing or nonverbal means. A strategic choice is to generate entailed acts only nonverbally, only verbally, using both modalities, or in random or specified variations. We also performed ranking experiments using TRAINS information-seeking dialogues where all communicative functions of a segment were considered: independent, entailed, implicated, and Turn Management default and side-effect functions.

Five TRAINS dialogues were (arbitrarily) selected containing 351 functional segments. All dialogues were annotated by three expert annotators. Table 8.11 presents the relative frequency of independent and implied communicative functions across dimensions in this data.

Table 8.11: Distribution of different types of communicative functions across dimensions for TRAINS dialogue in (%).

Dimension	Frequency					
	total	independent	entailed	implicated	default	side-effects
Task	37.6	100.0	0.0	0.0	0.0	0.0
Auto-Feedback	44.7	64.3	27.4	8.3	0.0	0.0
Allo-Feedback	9.1	21.9	78.1	0.0	0.0	0.0
Turn Management	44.7	56.7	0.0	0.0	21.7	21.6
Time Management	13.4	100.0	0.0	0.0	0.0	0.0
Contact Management	2.8	0.0	50.0	50.0	0.0	0.0
Discourse Structuring	5.1	83.3	0.0	16.7	0.0	0.0
Own Comm. M.	6.0	100.0	0.0	0.0	0.0	0.0
Partner Comm. M.	0.6	100.0	0.0	0.0	0.0	0.0
Social Obligation M.	2.0	100.0	0.0	0.0	0.0	0.0

The experiment was performed by untrained annotators: five undergraduate students who followed a course on pragmatics during which they were exposed to approximately four hours of lecturing on DIT, and participated in a few small-scale annotation exercises.
Pairwise inter-annotator agreement between the raters in terms of Cohen's kappa (Cohen, 1960) was measured for ranks 1, 2, 3 and 4, since at most four functions were assigned to segments. Table 8.12 presents the results. The kappa values range from 0.38 and 0.87, inter-rater agreement ranging from moderate (0.49) to near perfect (0.82).

Raters reached near perfect agreement on the most important function (rank 1). For rank 2, substantial agreement was reached. For ranks 3 and 4 the agreement was moderate.

Table 8.12: Cohen's kappa scores for communicative function ranking experiment per pair of raters.

Annotator pair	Rank 1	Rank 2	Rank 3	Rank 4
A vs D	0.87	0.76	0.54	0.5
A vs H	0.84	0.73	0.54	0.52
A vs K	0.81	0.67	0.55	0.38
A vs L	0.82	0.61	0.64	0.5
D vs H	0.8	0.66	0.34	0.41
D vs K	0.81	0.75	0.44	0.38
D vs L	0.83	0.55	0.36	0.41
H vs K	0.78	0.66	0.55	0.49
H vs L	0.82	0.66	0.54	0.6
L vs K	0.79	0.55	0.5	0.71
overall	0.82	0.66	0.5	0.49

Table 8.13 gives an overview of the average ranks assigned to communicative functions in different dimensions, when considering dimension combinations. Functions in the Task dimension are seen as the most important.

Along with a Task function, positive Auto-Feedback is generally seen as very important, and ranked higher than functions in other dimensions except if Discourse Structuring and Social Obligation acts are performed. In the latter situations, positive Auto-Feedback functions are not independent but implied (e.g. a reactive Greeting entails positive feedback). Positive Auto-Feedback is also ranked lower in combinations with Turn Management functions, when the speaker signals his intention to continue in the speaker role using a discourse marker, like *'and'*, *'then*, *'so'*.

Partner Communication Management and Own Communication Management are both concerned with problems or mistakes in speech production, whose resolution is, obviously, of great importance to dialogue participants. The same explanation is valid for the high score of negative Auto-feedback. Contact Management acts may be motivated by uncertainty about the partner's presence, which is an important aspect in telephone conversations, like the TRAINS dialogues.

From the results we conclude that:

- negative feedback acts but also acts concerning Own and Partner Communication Management have higher priority than the other acts that they may occur in combination;
- independent functions are more important than implied ones;
- task acts have higher priority than acts in other dimensions (this may be specific for task-oriented dialogues);
- independent and implied acts ensuring and maintaining contact between participants and structuring the dialogue have high priority (this may be specific for telephone dialogues);
- acts performed for managing time and turn are equally important.

Table 8.13: Average ranking of communicative functions in different dimensions and in different dimension combinations.

Dimension combinations	Task	Auto-F.	Allo-F.	TurnM.	TimeM.	ContactM.	DS	OCM	PCM	SOM
Task + Pos.Auto-F.	1.1	1.8								
Task + Pos.Auto-F.+ TurnM.	1.0	2.0		2.5						
Task + Turn	1.0			2.0						
Task + Pos.Auto-F.+ Pos.Allo-F.	1.0	2.0	2.5							
Task + OCM	1.1							1.6		
Task + TurnM.+ OCM	1.2			2.5				1.9		
Task + Pos.Auto-F.+ TimeM.	1.2	2.4			1.8					
Task + Pos.Auto-F.+ Pos.Allo-F.+ TurnM.	1.0	2.0	2.4	2.6						
Task + TurnM. + TimeM.	1.0			2.2	2.8					
Pos.Auto-F. + Pos.Allo-F.		1.1	1.8							
Pos.Auto-F. + Pos.Allo-F. + PCM		1.8	2.2						1.0	
Pos.Auto-F. + Pos.Allo-F. + TurnM.		1.1	1.9	1.7						
Pos.Auto-F. + Pos.Allo-F. + TurnM. + TimeM.		1.0	2.0	3.0	3.0					
Pos.Auto-F. + DS		1.7					1.2			
Pos.Auto-F. + DS + SOM		2.0					1.4			1.6
Pos.Auto-F. + SOM		2.0								1.0
Pos.Auto-F. + TurnM.		1.7		1.2						
Pos.Auto-F. + TurnM. + TimeM.		1.7		1.7	2.1					
Pos.Auto-F. + TurnM. + ContactM.+DS		2.4		2.6		2.2	1.8			
Neg.Auto-F. + TurnM. +TimeM.		1.7		1.7	2.1					
TurnM. + TimeM.				1.2	1.7					
TurnM. + OCM				1.6				1.3		
TurnM. + DS				1.7			1.2			
ContactM. + DS + SOM						1.7	2.4	1.9		

The results outlined here offer new insights into the multifunctionality of dialogue units. We identified a number of recurring patters which govern the perceived relative importance of communicative functions. Some of these patterns are related to the dialogue domain and settings. Others reflect general characteristics and constraints of the communication process. Still other constraints are brought about by semantic relations such as entailment or implicature.

8.4.3 Defining dialogue strategies

Additional conditions for how to deal with alternative possible dialogue acts may be motivated by particular dialogue settings, and offers a way to implement different dialogue strategies and styles of communication. Such options are mainly concerned with the following phenomena: (1) implicit *vs* explicit generation of certain entailed, implicated, default and side-effect functions; (2) attending to social obligations in dialogue; and (3) the choice between different combinations of modalities. In the next subsections we consider some cases where the generation of multiple dialogue act candidates offers such options.

Positive feedback

A strategic issue is whether or not to explicitly produce a dialogue act that is already implied by another one. In the case of a spoken dialogue system, giving explicit feedback can be a good strategy in view of the errors made in automatic speech recognition.

We investigated how much linguistic positive feedback is actually provided by humans in different types of dialogue. The upper part of Table 8.14 presents the distribution of positive auto-feedback acts in different corpus data. It can be observed that explicit linguistic positive auto-feedback was provided more often in MapTask dialogues when the participants have no direct visual contact (93.4%) compared when they do have visual contact but cannot see task-related actions (77.8%), and even less often in AMI meetings where the participants are involved in face-to-face interaction and are able to observe all movements and actions (59.4%). We showed in Chapter 6 Section 6.2 that when participants have access to all modalities and can observe all of each other's actions, auto-feedback is provided through non-verbal displays and signals. In the TRAINS dialogues explicit linguistic positive auto-feedback occurs again relatively often (79.8%); this is not surprising since these dialogues are conducted over the telephone.

Similar considerations apply to allo-feedback. Consider the following example:[10]

(92) U1: How far is it from Avon to Danville?
S1: Three hours
U2: Three hours, that's right

In U2 several acts are performed: (1) an explicit positive auto-feedback act at the level of perception; (2) a positive auto-feedback at the levels of execution (S1 is evaluated and accepted); and (3) an entailed positive allo-feedback act concerning S's understanding of the question U1.

Table 8.14 shows that this type of feedback occurs more often when the participants have visual contact. This can perhaps be explained by the fact that partners who see each other's actions and movements tend to comment on those.

[10] From the TRAINS dialogue corpus.

Table 8.14: Relative frequencies of linguistic positive feedback acts in various types of human-human dialogues (in %).

Type of positive feedback	AMI	TRAINS	MapTask	
			eye-contact	no eye-contact
auto-feedback	22.1	57.8	38.0	55.5
of this				
explicit independent	57.9	76.4	73.8	88.5
explicit implied	7.5	3.4	4.0	4.9
implicit implied	34.6	20.2	22.2	6.6
allo-feedback	14.5	8.8	13.6	8.6
of this				
explicit independent	1.8	19.8	28.6	28.1
explicit implied	39.5	32.6	14.3	6.3
implicit implied	58.7	48.4	57.1	65.6

A general conclusion that we can draw with respect to positive allo-feedback is that a safe strategy would be not to generate such acts in explicit linguistic form. For the use of other modalities we make suggestions later in this section.

Turn Management

Another strategic consideration is whether to generate explicit turn management acts such as Turn Take, Turn Accept, Turn Release and Turn Assign. In Section 8.3 it was noticed that this heavily depends on the dialogue setting. For example, in high risk dialogues where participants are under severe pressure to perform accurate actions in short time, such as military settings, air traffic control dialogues, health and public safety emergencies, or crisis handling dialogues, it can be of crucial importance to make clear who is the current and intended next speaker, as in the following example:[11]

(93) FO: steel one niner this is gator niner one adjust fire polar **over**
FDC: gator nine one this is steel one nine adjust fire polar **out**
FO: direction five niner four zero distance four eight zero **over**
FDC: direction five nine four zero distance four eight zero **out**

Social obligations and politeness

The role of social obligations management acts in the process of selecting and combining dialogue acts is also a matter of strategy and choosing an appropriate style of communication.

Social behaviour that dialogue partners are expected to exhibit as being cooperative, rational and social agents is not limited to the performance of appropriate social obligations acts, but also relates to the issue of whether to generate direct or indirect dialogue acts. Indirect acts, like conditional requests, are often perceived as more polite since they offer the addressee a way to refuse to perform an action without 'losing face'. Similarly, statements that contain subtle expressions of uncertainty are perceived as less assertive or offensive.

[11] From the Radiobot-CFF dialogue corpus provided in (Roque et al., 2006).

Table 8.15: Forms of functional segments for dialogue acts with general-purpose functions (frequency in %).

GP function	One-token	Phrase	Clause	Sentence
Set Question	0.0	0.0	0.0	100.0
Choice Question	0.0	0.0	0.0	100.0
Propositional Question	0.0	16.7	5.6	77.7
Check Question	0.0	38.9	0.0	61.1
Inform	0.0	0.0	30.0	70.0
(Dis-)Agreement	100.0	0.0	0.0	0.0
Set Answer	0.0	16.7	0.0	83.3
Propositional Answer	81.5	7.4	0.0	11.1
(Dis-) Confirm	92.3	0.0	0.0	7.7
Request	0.0	11.1	0.0	88.9
Accept Request	66.7	0.0	0.0	33.3
Suggest	0.0	30.0	30.0	40.0

8.4.4 Linguistic constraints on dialogue act combinations

Some combinations of dialogue acts may not carry any logical or pragmatic conflicts, but may be hard to express simultaneously in natural language. Linguistic constraints can thus be viewed as an additional filter on dialogue act combinations in speech-only (or text-only) dialogue systems, and as constraints on the use of language in a multimodal system.

Linguistic constraints may be theoretical in nature, based on consideration of grammatical well-formedness, or empirical, based on the observed use of dialogue act combinations. We only consider empirically-based constraints here. Empirical linguistic constraints can be determined (1) through the analysis of dialogue act co-occurrences; and (2) by similarity and distance measurements between functional segments represented by feature vectors.

The unit of expression that we mainly considered so far is the functional segment. The kinds of functional segment that occur in dialogue can be divided into (1) *one-token segments*, e.g. inarticulate feedback, stallings, turn management acts; (2) *token sequences* that do not form grammatical units; (3) *phrases*, *clauses*, and *sentences*, that often form functional segments with a general-purpose function. Theoretical linguistic constraints apply only to the latter form of segments. Empirical linguistic constraints may apply to all forms of segments as well as to larger units: utterances, which may consist of several functional segments of various form, and turn units.

To gain insight into the use of dialogue acts combinations, we performed small-scale experiments comparing the linguistic properties of functional segments in AMI corpus data.

Combining dialogue acts with general-purpose functions

Functional segments with general-purpose functions are mostly sentences, clauses or phrases; one-token segments are rare and occur only for Propositional Answer, Confirmation, Agreement and Accept Request (see Table 8.15).

For testing the linguistic similarities and differences of phrases, clauses and sentences, we represented annotated segments by vectors with 9 prosodic values (duration, min, max, mean, standard deviation in pitch, pitch slope, fraction voiced/unvoiced frames, voice breaks and intensity) and 1623 values for word tokens occurring in all segments. Tokens were weighted

Table 8.16: Lexical similarity of segment vectors for different general-purpose communicative functions.

GP function	Inform	Instruct	Suggest	SetAnswer	SetQuestion	Prop.Question
Inform	1.000					
Request	.807	1.000				
Suggest	.869	.756	1.000			
SetAnswer	.851	.674	.751	1.000		
SetQuestion	.845	.764	.809	.698	1.000	
Prop.Answer	.391	.235	.270	.398	.279	
Prop.Question	.235	.149	.190	.232	.258	1.000

according to the *tf/idf* function, i.e. the frequency of occurrence of a token in segments with a certain communicative function was multiplied by the inverse communicative function frequency. This weighting was done for all tokens, for the first and last segment tokens, and for bigrams. We then measured semantic similarity between vectors for segments that have different general purpose function, by using the following formula, where *w* stands for weight, *t* for token and *v* for vector:

$$sim(\overrightarrow{v_j}, \overrightarrow{v_k}) = \frac{\sum w_{t,v_j} \times w_{t,v_k}}{\sqrt{\sum w_{t,v_j}^2} \times \sqrt{\sum w_{t,v_k}^2}}$$

Table 8.16 presents the results of calculating the similarity of the token vectors of segments that have different general-purpose functions, only the most frequent functions were considered. Lexically, Suggest is close to Inform, and Inform is relatively close to SetAnswer; they share more or less the same vocabulary. Propositional and Set Question, by contrast, are rather different lexically. In fact, Propositional Questions differ lexically from all other tested dialogue acts. The same was observed for Propositional Answers.[12]

We also assess the prosodic differences between segments with different general purpose functions, measuring distances between prosodic segment vectors pair-wise using Euclidean distance (see Chapter 3 for calculations). Table 8.17 presents the results. Note e.g. that Inform is prosodically close to Set Answer and to Suggest, but quite distant from Request, Set Question and even more from Propositional Question. Request and Suggest, by contrast, are prosodically relatively close to Set Questions, and they are similar in intensity contour and speaking rate. As was to be expected, answers clearly differ prosodically from questions.

It can be concluded here that it is rather infelicitous to combine Answers and Inform with Questions, or generally forward-looking general-purpose functions with backward-looking general-purpose functions; Instructs and Requests with Suggestions, i.e. two or more directives that differ in strength of the pressure put on the addressee to perform a certain action, e.g. it is questionable whether it is possible to combine an Instruct (*'Place this button here'*), Request (*'Can you place this button here'*) and/or Suggest (*'You may place this button here'*).

In our data general-purpose functions never co-occur. Although some combinations are possible, in view of their lexical and prosodic properties, they are highly unusual. If two dialogue acts with general-purpose functions are selected as candidates for generation, it would therefore seem advisable not to generate them in one segment, but rather in two segments as

[12]Note that in our data the majority of Propositional Answers were one-token segments in the form of 'yes' or 'no' answers.

Table 8.17: Prosodic distances of segment vectors for different general-purpose functions.

GP function	Inform	Instruct	Suggest	SetAnswer	SetQuestion	Prop.Question
Inform	.000					
Instruct	51.823	.000				
Suggest	6.292	48.386	.000			
SetAnswer	8.829	43.965	9.162	.000		
SetQuestion	25.832	28.690	21.127	19.731	.000	
Prop.Answer	57.557	15.502	53.710	49.635	33.971	
Prop.Question	48.130	16.019	43.295	41.454	22.390	.000

Table 8.18: Frequency of occurrence of dialogue acts with general-purpose functions at specified position.

GP function	Single segment turn units	Multi-segment turn units		
		1st segment	middle	last segment
Set Question	20.0	0.0	20.0	60.0
Choice Question	0.0	0.0	0.0	100.0
Propositional Question	22.2	16.7	33.3	27.8
Check Question	21.1	10.5	26.3	42.1
Inform	16.8	7.4	51.0	24.8
(Dis-)Agreement	55.6	27.8	11.1	5.6
Set Answer	16.7	33.3	50.0	0.0
Propositional Answer	57.1	14.3	21.4	7.4
(Dis-) Confirm	84.6	15.4	0.0	0.0
Request/Instruct	18.2	9.1	18.2	54.5
Accept Request	33.3	66.7	0.0	0.0
Suggest	20.0	0.0	40.0	40.0

part of a turn unit (or generate only one of them). The question then arises in what order the two segments should be generated. To answer this question, we analysed the construction of turn units in AMI data (see Table 8.18 and found out that there are certain general patterns in participants behaviour in this respect. Questions and directives such as Request, Instruct and Suggestions tend to occur at the last position in turn unit, unless they are rhetorically related as antecedents to subsequent segments, e.g. questions followed by informs that elaborate, clarify or justify these questions. Answers, and acceptances of a request mostly occur in single segment turn units or as first segments in multi-segment units, and very rare close a turn unit. Inform segments may form a turn unit on their own, but most of the time occur in the middle of a turn unit, and sometimes close it.

Combining dialogue acts with general-purpose and dimension-specific functions

Table 8.19 shows the co-occurrence of general-purpose and dimension-specific functions. Some combinations occur frequently in our data. Speakers often signal that they want to have the turn while performing a dialogue act with a general-purpose function, e.g. by using a discourse marker for this purpose (see Chapter 6, Section 6.4). Some combinations do not occur at all; for example, Inform and responsive information-transfer functions do not co-occur with feedback elicitation, since elicitations are linguistically similar to questions.

Table 8.19: Co-occurrence of general-purpose and dimension-specific communicative functions in a single segment.

DS \ GP	Inform	Agreement	Disagreement	SetA	PropA	Confirm	Check	SetQ	PropQ	Request	Acc.Req	Suggest
Auto-Feedback:												
pos.evaluation :	0.2	0.7	0.0	0.0	0.0	0.0	3.6	0.0	0.0	0.0	0.0	0.0
pos.execution:	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	43.2	4.1
pos.interpretation:	0.0	91.8	0.0	21.4	43.5	90.0	1.0	1.1	8.5	0.0	0.0	0.0
neg.evaluation :	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.9	8.1	0.0	0.0
neg.execution :	0.2	0.0	3.0	0.0	0.0	0.0	0.5	0.0	0.9	0.0	0.0	0.0
Allo-Feedback:												
neg.execution:	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
elicitation :	0.0	0.0	0.0	0.0	0.0	0.0	7.6	0.0	10.7	0.0	0.0	0.0
TurnM.:												
grab :	1.6	8.9	6.1	0.8	20.0	13.3	4.6	3.4	1.7	0.0	2.7	2.1
take :	0.3	69.9	0.0	3.2	11.8	66.7	0.0	0.0	0.2	0.0	8.1	0.5
acc :	0.0	0.0	0.0	0.0	1.2	3.3	0.0	0.0	0.0	0.0	21.6	0.0
keep : .	9.5	11.6	18.2	9.5	15.3	6.7	9.1	5.7	12.0	3.5	13.5	11.4
release :	0.0	68.5	0.0	6.3	11.8	80.0	2.0	10.3	2.1	1.2	16.2	0.0
assign :	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.4	2.4	8.1	8.1	0.0
TimeM.:												
stall : .	2.8	0.7	9.1	4.0	1.2	0.0	2.0	1.1	3.2	0.0	2.7	2.6
pausing : .	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DS:												
introduction :	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
shift :	0.7	1.4	0.0	0.0	0.0	0.0	7.6	0.0	0.2	2.3	0.0	0.0
change-announce :	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.3	8.1	0.0	0.0
pre-close :	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 8.20 shows the frequency of occurrence of communicative functions which do not co-occur in a single segment, in segments at different positions within a turn unit. Feedback acts are apparently better generated at the first position in a multi-segment turn unit, possibly together with turn-initial acts and stallings (see next subsection). Own communication acts occur in the middle or close to the end of a turn unit; Partner Communication acts, by contrast, are preferably generated first and may be followed by other acts, e.g. often by elaborations and feedback acts. Contact and Social Obligation Management acts do not occur in segment sequences where one of the dialogue acts has a general-purpose function.

Table 8.20: Frequency of occurrence of dialogue acts with dimension-specific functions at specified position if a turn unit has at least one dialogue act with a general-purpose function.

DS function	Multi-segment turn units		
	1st segment	middle	last segment
Positive Auto-Feedback	88.7	6.3	5.0
Negative Auto-Feedback	66.7	0.0	33.3
Allo-Feedback	50.0	33.3	16.7
Turn-initial	100.0	0.0	0.0
Turn-keep	0.0	99.7	0.3
Turn-final	0.0	6.7	93.3
Time Management	28.7	70.7	0.6
DS	35.2	47.1	23.5
OCM	0.0	91.1	8.9
PCM	100.0	0.0	0.0

Combining dialogue acts with dimension-specific functions

Table 8.21 shows the syntactic forms of segments that express dialogue acts with dimension-specific functions. Since such acts do not have an articulate semantic content, these segments are mostly linguistically simple; the majority are one-token segments. Exceptions are social obligation acts, e.g. Self-Introduction, which often have the form of a full sentence.

Table 8.21: Forms of functional segments for dialogue acts with dimension-specific functions (frequency in %).

DS function	One-token	Phrase	Clause	Sentence
Auto-Feedback	93.1	6.9	0.0	0.0
Allo-Feedback	75.0	25.0	0.0	0.0
Turn Management	100.0	0.0	0.0	0.0
Time Management	98.6	0.0	0.0	1.4
Contact Management	100.0	0.0	0.0	0.0
Discourse Structuring	100.0	0.0	0.0	0.0
OCM	83.3	16.7	0.0	0.0
PCM	81.5	17.5	0.0	0.0
SOM	28.6	0.0	0.0	71.4

Table 8.22 shows the co-occurrence patterns of dialogue acts with dimension-specific functions in a single segment, as observed in data. Time Management acts are linguistically compatible with all other acts, since stallings can be realized by, for example, lengthening of a word

or syllable. Similarly, Turn Management acts are not in conflict linguistically with other acts. Auto-Feedback acts for instance often have a turn-initial function (except for backchannels), while Allo-Feedback elicitations have a turn-final function signalled by a rising intonation. Such acts can be expressed in a single functional segment by using features such as speaking rate (slowing down to stall for time), intonation (rising to give a turn away and to elicit feedback), or intensity (louder voice to claim or to keep the turn).

Table 8.22: Co-occurrences of communicative functions across dimensions in a single segment, expressed in relative frequency in %. (Read as follows: percentage of segments having a function in the dimension of the column, which also has a function in the dimension of the row.)

	Auto-F.	Allo-F.	Turn M.	Time M.	CM	DS	OCM	PCM	SOM
Auto-F.	0.0	0.0	11.8	0.9	0.0	10.7	2.0	0.0	14.3
Allo-F.	0.0	0.0	46.3	0.3	8.8	1.0	0.0	1.0	0.0
TurnM.	48.8	45.1	0.0	17.5	25.0	14.6	67.3	25.8	14.3
TimeM.	5.5	18.2	54.4	0.0	6.2	0.8	2.0	16.1	3.2
CM	0.2	0.1	0.0	0.6	0.0	5.6	0.0	0.0	2.0
DS	0.3	0.1	1.3	0.2	27.1	0.0	0.0	0.0	8.2
OCM	0.8	0.9	3.1	0.2	0.0	2.2	0.0	0.0	0.0
PCM	0.0	0.1	0.0	0.1	0.0	0.7	0.0	0.0	0.0
SOM	0.1	0.0	0.1	0.0	15.3	0.3	0.2	0.0	0.0

Table 8.23 shows the occurrence of communicative functions which do not co-occur in a single segment, but in multi-segment turn units. For example, it seems preferable to report about speaker's and partner's processing in separate segments, unless one implies the other. The ordering patterns of segment sequences with dimension-specific functions can be used for the generation of such turn units.

Table 8.23: Frequency of occurrence of dialogue acts with dimension-specific functions at specified position in multi-segment turn units.

	Single segment	Multi-segment turn units		
	turn units	1st segment	middle	last segment
Positive Auto-Feedback	70.0	21.3	6.0	2.7
Negative Auto-Feedback	50.0	16.7	0.0	33.3
Allo-Feedback	100.0	0.0	0.0	0.0
Allo-Feedback (elicitation)	25.0	25.0	50.0	0.0
Turn-initial	4.7	95.6	0.0	0.0
Turn-keep	0.0	35.6	64.0	0.4
Turn-final	82.6	8.7	0.0	8.7
Time Management	0.6	28.1	70.7	0.6
DS	0.0	44.4	33.3	22.2
Contact Management	0.0	100.0	0.0	0.0
OCM	0.0	0.0	85.0	15.0
PCM	100.0	0.0	0.0	0.0
SOM	14.3	42.9	28.5	14.3

In sum, we showed how to define possible linguistic constraints on dialogue act combinations. The results presented here are based on corpus data that is obviously limited, and may

be specific for the setting of these dialogues. The conclusions based on these results should be considered as recommendations rather than generic rules, and most importantly as indicative for the kind of conclusions that can be reached for a particular type of dialogue in a particular setting, based on empirical co-occurrence data.

8.5 Conclusions

In this chapter we have, first, discussed a multidimensional computational context update model. Since an utterance, which can be multifunctional, when understood by a dialogue participant evokes certain changes in the participant's context model, and these changes typically do not affect the entire context model, a context model should be structured in such a way that certain parts can be updated independently while others remain unaffected. We have discussed the main components of a multidimensional context model and the formalisation of update effects. We showed how the model is updated by the application of a few general mechanisms which reflect the assumed cooperativity, rationality, and sociality of dialogue participants and what effects this has on the information states of dialogue participants.

We proposed a context-driven approach to dialogue generation based on the expected understanding and adopting effects that each dialogue act creates. These effects that can be applied in a given context are reliable tools for constructing speaker's plans for dialogue continuation in the form of multiple dialogue acts that are candidates for being expressed in multifunctional utterances.

The multidimensional view on dialogue modelling suggests that in generating dialogue behaviour, dialogue acts may be selected from different dimensions simultaneously and independently, and then combined into multifunctional utterances. This procedure can be formulated as consisting of several subprocesses where:

- dialogue act candidates are inspected for any logical and pragmatic conflicts, which may be resolved by cancelling or postponing lower-priority acts.
- dialogue act candidates are ordered according to their relative importance given the global and local dialogue context, resulting in cancelling or postponing some dialogue act candidates.
- the remaining list of partially ordered candidates is evaluated from a pragmatic and dialogue strategic point of view, again possibly resulting in cancelling or postponing some of the dialogue act candidates.
- combinations of dialogue acts are selected that can actually be realised in a multifunctional segment or sequences of segments, taking the available modalities (other than speech) into account.

The multidimensional approach to dialogue act interpretation and generation as outlined in this chapter contributes to the design of the generation part of a Dialogue Manager. Considering various dimensions simultaneously contributes not only to more accurate and adequate interpretation of a user's dialogue behaviour, but also to the generation of interactive behaviour that is natural to human and exploits the full potential of spoken and multimodal interaction. It enables the generation of utterances that are multifunctional by design, addressing several aspects of the communication at the same time to make the interaction more effective and efficient.

Chapter 9

Conclusions and perspectives

In this chapter we formulate the main conclusions of the research reported in this thesis, and indicate perspectives and directions for future research that builds on this work.

9.1 Conclusions

Dimensions of communication Communication is a complex, multi-faceted activity. Utterances in dialogue often have multiple communicative functions which must be taken into account in order to avoid errors and misunderstandings, and to support a dialogue that is effective and efficient. In order to describe and model the multifunctionality of dialogue contributions it is helpful to analyse the functions that an utterance may have in multiple dimensions. No clear definition of 'dimension' has been proposed in the literature, however. We analyzed a range of approaches that use the notion of dimension (Allen and Core, 1997; Larsson, 1998; Soria and Pirrelli, 2003; Popescu-Belis, 2004) and pointed out that the term 'dimension' is used with different senses, each of them unsatisfactory in several respects. We have provided a conceptually clear definition of 'dimension', and have put forward five criteria which a set of dimensions should meet: theoretical justification, empirical validity, orthogonality, reliable recognisability, and compatibility with existing annotation schemes where possible.

Analysing a variety of theoretical work on dialogue analysis and modelling, applying a range of tests to annotated dialogue corpora, and taking 18 existing annotation schemes into account, ten dimensions are identified which are shown to meet these criteria: Task, Auto-Feedback, Allo-Feedback, Turn Management, Social Obligations Management, Own Communication Management, Discourse Structuring, Partner Communication Management, Time Management and Contact Management.

This set of dimensions has been adopted in the latest release (Release 5) of the DIT $^{++}$ taxonomy and in the annotation scheme developed in the LIRICS project. The results of this study have also been the basis for choosing the nine dimensions of the ISO dialogue act annotation standard 24617-2.

Communicative functions The assignment of meanings to units in dialogue in terms of communicative functions presupposes a way to segment a dialogue into meaningful units. We identified various types of dialogue units that play an important role in dialogue analysis and modelling: turn units, utterances, functional segments and discourse units. For each type of

unit, we discussed the role and purpose in dialogue analysis and we identified certain structural and semantic relations that may connect them.

We pointed out that existing dialogue act taxonomies fail to capture nuances in the performance of communicative actions relating to uncertainty, conditionality and sentiment. Participants in a dialogue do not just exchange messages by simple statements, clear-cut answers and direct requests. They may be less straightforward in expressing their communicative intentions, formulating a question indirectly or accepting a request conditionally. They often indicate their attitude toward communicative partners, toward what is said, or toward things that may be done. We developed a way to deal with such phenomena in terms of 'qualifiers' that can be attached to a communicative function, in order to describe the speaker's intentions more accurately. The proposed qualifiers for dealing with uncertainty, conditionality and sentiment have been adopted in the ISO 24617-2 standard.

In order to obtain an adequate and empirically valid characterization of the multiple functionality of dialogue contributions, we analysed the forms of multifunctionality that occur in dialogue data. It was argued that a good understanding of the nature of the relations among the various multiple functions that a dialogue unit may have, and how these units relate to other units in dialogue, is a prerequisite for defining a computational update semantics for dialogue utterances. We presented an empirical account and analytical examination of forms of multifunctionality in dialogue units of various sorts and their relation to dimensions of communication. The various functions of a unit in dialogue are either independent, and occur by virtue of a local and contextual features, or because certain semantic relations exist between the communicative functions of functional segments. The latter occurs when functions have entailment relations, conversational implicatures, default functions, or side-effects.

We have shown how the (multi-)functionality of dialogue units can be recognized based on observable behavioural features in data-oriented way. A token-based approach combining the use of local classifiers, which exploit local utterance features, and global classifiers which use the outputs of local classifiers applied to previous and subsequent tokens, is shown to result in excellent dialogue act recognition scores for unsegmented spoken dialogue.

Features of dialogue utterances A requirement for distinguishing a communicative function is that there are ways in which a sender can indicate that his behaviour should be understood as having that particular function, shaping his behaviour so as to have certain observable features which are indicative for that function in the context in which the behaviour occurs. This requirement puts all communicative functions on an empirical basis.

We have presented a detailed analysis of how dialogue participants express the intended functions of their dialogue contributions, and how they recognise the intended functionality of partner utterances. We have focused in particular on interaction management acts, since they form a relatively large part of what happens in natural conversation and are largely responsible for the naturalness and smoothness of spontaneous dialogue. They have however, largely escaped a detailed analysis and resisted an integrated formal account. We identified features from the physical realisation of a dialogue utterance in context that can help to predict what type of dialogue act is performed. The properties of the most frequently occurring types of acts in our data were analysed, such as feedback acts, turn management acts, and discourse structuring acts. The relevant features are not restricted to language-related properties of utterances, but include nonverbal aspects as well. We reported results of explorative studies, observations from annotated data, statistical analyses, and perceptual experiments. One of the conclusions is that a well-worked out, fine-grained, open multidimensional dialogue act taxonomy such as DIT^{++} (but also other multidimensional taxonomies like DAMSL, MRDA or Coconut) is

suitable for this purpose when some adjustments are made in order to deal with the uncertainty and sentiment that is expressed by nonverbal modalities.

Dialogue context properties In order to understand what happens in dialogue it is not sufficient to consider the content and function of its segments in isolation. We argued that the recognition of communicative actions should be based on the understanding of coherent discourse, not just of understanding independent actions. Successful interpretation of communicative acts in dialogue and their generation is dependent on global and local context properties. Dialogue acts are often semantically dependent on one or more dialogue acts that occurred earlier in the dialogue, in the sense that their semantic content can only be determined by taking the semantic content of these preceding dialogue acts into account. Local context properties are essential for successful automatic dialogue act recognition. Global context properties such as type, domain of dialogue, its general settings and participant's knowledge and assumptions about each other heavily influence how the meaning of utterances can be recognized and how dialogue acts can be selected for generation.

Dialogue context models provide the basis for interpreting the speaker's behaviour and for decisions about future actions. An important issue is therefore what kinds of information should be included in a participant's context model. A dialogue context model originally proposed in (Bunt, 1994) is structured into five components: (1) the participant's information about the underlying task and its domain ('Semantic Context'); (2) the participant's state of processing ('Cognitive Context'); (3) the availability and properties of communicative and perceptual channels, and the partner's presence and attention ('Physical/Perceptual Context'); (4) communicative obligations and constraints ('Social Context'); and (5) the preceding dialogue contributions and possible discourse plans ('Linguistic Context'). A dialogue segment, when understood by a dialogue participant as a dialogue act with a certain communicative function and semantic content, evokes certain changes in the participant's context model. These changes typically do not affect the entire context model, but only certain parts of it. Which part of a context model is affected by a dialogue act depends on the type of its semantic content. Given the formalization of updates on the dialogue context model that was proposed, it was shown how context motivates and enables communicative actions, how the information states of dialogue participants undergo certain changes when they understand the corresponding dialogue behaviour, and how information is transferred from one dialogue participant to another. It was demonstrated how such update effects together with general principles of cooperativity and rationality give rise to the generation of dialogue acts in multiple dimensions.

General conclusions Coming to more general conclusions, a first conclusion is that the systematic application of a multidimensional view on communication in combination with modelling the relevant types of information in a structured representation of dialogue context leads to a better understanding of human dialogue behaviour and enables better computational modelling of multimodal dialogue. We showed that the multifunctionality of dialogue contributions is a phenomenon that cannot be ignored, and that should rather be exploited. The complexity of multifunctional dialogue behaviour has caused scepticism as to whether it is computationally possible or attractive to develop dialogue models which can deal with multifunctionality. We argued, however, that many problems can be solved when using a multidimensional approach to dialogue analysis and modelling, based on well-defined concepts of 'dimension' and 'dialogue act' supported by detailed empirical analysis of dialogue behaviour. The obtained insights in the nature, forms and linguistic and non-linguistic manifestation of multifunctionality opens the perspective of a dialogue system that understands and generates utterances which are multifunctional, by design.

A second general conclusion is that the use of fundamental concepts and insights from dialogue theory is generally useful for an adequate analysis of human dialogue behaviour, for modelling this behaviour, and for the design of dialogue systems. In particular, it may be observed that context-driven dialogue understanding and generation make use of assumptions concerning cooperativity, rationality and sociality of dialogue participants behaviour, showing that such assumptions are useful, if not indispensable in computational modelling of dialogue.

A third general conclusion is that the analytical and empirical studies reported in this thesis have contributed to the advancement of the state of the art in dialogue annotation. The detailed investigation of spoken and multimodal dialogue, of the semantics of functional segments, and of semantic and structural relations between them, has contributed to the specification of the new ISO standard 24617-2 for dialogue act annotation, and, hand in hand with that, the development of the latest release (Release 5) of the DIT^{++} annotation scheme as a strictly compatible extension of that standard.

To summarise, the main ideas, concepts and assumptions of a theory of dialogue, such as the 'information-state' theory in general and Dynamic Interpretation Theory in particular, that consider the meaning of communicative behaviour in terms of the changes in the participants' state of information upon successful communication, combined with a multidimensional view on dialogue communication, open the way to design effective and efficient dialogue systems that are flexible enough to exploit the full potential of spoken and multimodal interaction.

9.2 Perspectives and future directions

Besides having produced the results and conclusions discussed in the previous section, this thesis also raises new issues and suggests new opportunities and directions for future work that exploits the results of this thesis.

Dialogue act annotation and corpus construction Given the importance of annotated corpora for a wide range of linguistic applications, there is a need for annotation schemes that are populated with empirically as well as theoretically well-motivated concepts. For dialogue act annotation such concepts concern the definition of communicative functions, dialogue segmentation, and the definition of relations between dialogue segments; and in the case of multidimensional schemes also the definition of dimensions. The ISO 24617-2 standard annotation scheme and the DIT^{++} release 5 scheme, to which this thesis has contributed, are comprehensive, application-independent schemes whose concepts are indeed empirically and theoretically well-motivated, and may be exploited for constructing annotated dialogue corpora. Both the ISO and the DIT^{++} schemes cannot be expected to be ideal for every kind of dialogue analysis, for every task domain, for every kind of dialogue, and for every annotation purpose, but the general principles underlying the design of the schemes and the DiAML annotation language enable extensions, modifications, and restrictions of the schemes and the annotation language, as the need arises for particular applications. Future efforts can for instance be directed towards defining sub-taxonomies of domain-specific communicative functions, which can be plugged in in the ISO or DIT^{++} schemes for different applications and purposes.

An important goal when creating annotated language resources is their interoperability. The dialogue research community still does not have large amounts of annotated dialogue data at its disposal, compared to other linguistic communities. Moreover, the available resources are only partly compatible with each other. Many recently developed annotation languages are XML dialects, which enables data matching, search and application; it may be useful to

consider the possibility of using DiAML, also rooted in XML but equipped with a formal semantics, as an interlingua for converting between alternative representations.

Multimodal dialogue act recognition In Chapter 6 of this thesis we studied in detail the interpretation of nonverbal behaviour that deals with interaction management. We analysed three main aspects of that: feedback, turn taking and structuring the discourse. Certainly, nonverbal behaviour may addresses other dimensions. Other dialogue phenomena that deserve more detailed study in this respect include speech editing phenomena, mechanisms for establishing and maintaining contact, for managing time, for dealing with social obligations and constraints, and information status and affect.

A limitation of this thesis concerns the use of nonverbal features in machine-learning experiments. We do not have enough transcribed and annotated data of sufficient quality to incorporate non-verbal features into automatic recognition processes. To incorporate features from multiple modalities into classification experiments would be a challenging and interesting topic for future research.

Related to the previous point, it would be interesting to explore the use of high-level features obtained from other expert knowledge. We based our recognition tasks strictly on low-level features that are automatically extractable from the raw data, thereby eliminating errors that may occur at higher levels such as syntactic and semantic parsing. This decision is perfectly justified, but there is evidence from recent research that parsing techniques have advanced enough to enable us to obtain high-quality data annotated with syntactic information.

Finally, the performance of other machine-learning algorithms on the dialogue act recognition task would be worth to explore, e.g. machine learning techniques based on Conditional Random Fields that directly incorporate the interaction between local decisions and global decisions into the learning procedure.

Automatic utterance understanding The automatic, incremental recognition of communicative functions and dimensions in unsegmented spoken dialogue, reported in Chapter 7, can be seen as important step towards the automatic (incremental) understanding of dialogue utterances. Full utterance understanding calls for extending this work with (a) the recognition of rhetorical, functional dependence, and feedback dependence relations; and (b) the construction of a representation of the semantic contents of dialogue acts. The first of these can conceivably be achieved by applying the recognition approach, developed in Chapter 7, to sufficiently large corpora annotated with these relations. The second is more challenging, but we can see two directions that seem promising.

One direction is to integrate the incremental recognition of communicative functions with incremental syntactic and semantic parsing, and to exploit the interaction of syntactic, semantic and pragmatic hypotheses in order to understand incoming dialogue segments incrementally in an optimally efficient manner. We think that this is feasible when using a multidimensional approach to segmentation. Multidimensional segmentation solves a number of problems relating to "disfluencies" in speech, and leads to focus on constructing semantic representations for relatively simple and grammatically well-formed fragments of speech.

Another direction is to exploit the possibilities of extending dialogue act annotations with semantic annotations of other types, e.g. those marking up events, co-reference, semantic roles, time and location. According to the dialogue act-theoretical framework that we have used, the semantic content of a dialogue act is typically either an eventuality or a proposition. Adding this distinction to the metamodel proposed in Figure 4.9 would open the way for connecting with the metamodels used in ongoing ISO projects concerned with the annotation of time and events, space, and semantic roles, which could be very helpful for clarifying the relations be-

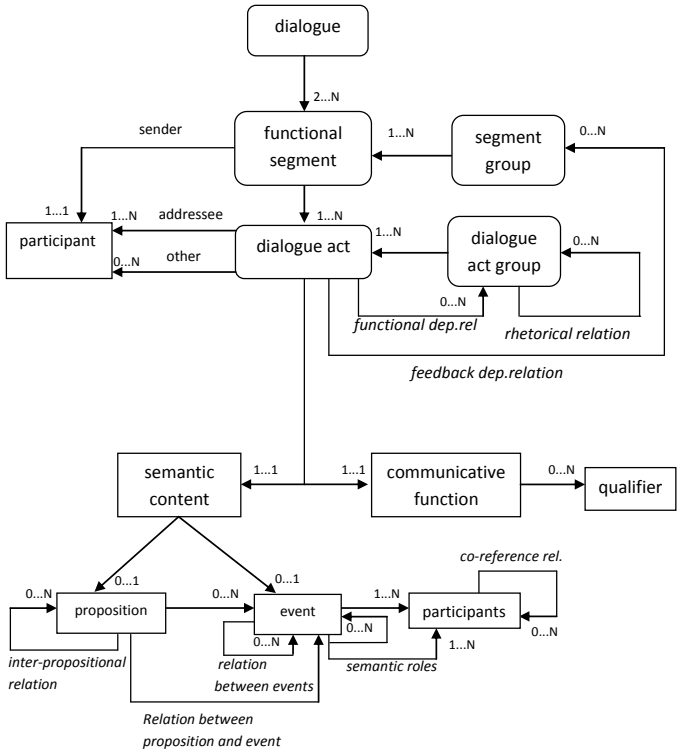


Figure 9.1: Extended metamodel of dialogue units, relations between them and other semantic entities.

tween the semantic phenomena targeted by these projects. Figure 9.1 may be considered as a step in this direction. Note that the annotations of events, locations, time, and semantic roles in these ISO projects have a formal semantics (see Bunt, 2007b; 2009c; Bunt and Overbeeke, 2008a; 2008b), so the more of these kinds of semantic annotations are added, the more information about semantic content becomes available. Again, an interesting challenge would be to do this in an incremental way.

Multidimensional dialogue management The detailed context model, described in Chapter 8, invites an implementation and evaluation in the setting of a dialogue system, together with the mechanisms for dialogue and context management that we also described. Initial implementation by Keizer and Bunt (2007) in the PARADIME dialogue manager, by Keizer and Morante (2007) in the DISCUS context update system, and by Petukhova et al. (2010) in the study of constraints on dialogue act combinations, suggest that this seems as interesting and feasible direction to go.

A particular challenge for future work would be the implementation of incremental context update semantics based on the interpretation of partial input. Following the incremental approach participants' information-states will be updated based on available partial input interpretation. These updates will be kept in the pending context and (incrementally!) evaluated for

consistency. If inconsistencies occur, this may also mean that initial interpretation is wrong and another hypothesis may be considered. This improves the quality of interpretation at earlier processing stages. If no inconsistencies occur, the context update process may go ahead and trigger (incrementally!) the generation of candidate dialogue acts. In this way an incremental Dialogue Manager could be designed, which takes care of deciding which action to take next in the dialogue generating dialogue acts in several dimensions simultaneously even before the user finishes his turn. One of the future tasks is to implement a complete dialogue system using the approach described in the thesis in order to evaluate the model and its components in a real dialogue setting.

Bibliography

[Afantenos and Asher, 2010] Afantenos, S. and Asher, N. (2010). Testing sdrt's right frontier. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1–9, Beijing, China.

[Ahn, 2001] Ahn, R. (2001). *Agents, Objects and Events. A computational approach to knowledge, observation, and communication. PhD Thesis*. Eindhoven University of Technology, The Netherlands.

[Aist et al., 2007] Aist, G., Allen, J., Campana, E., Gomez Gallo, C., Stoness, S., Swift, M., and Tanenhaus, M. K. (2007). Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. In Arstein, R. and Vieu, L., editors, *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, pages 149–154, Trento, Italy.

[Alexandersson et al., 1998] Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Meier, E., Reithinger, N., Schmitz, B., and Siegel, M. (1998). *Dialogue acts in Verbmobil-2*. DFKI Saarbrücken, University of Stuttgart; TU Berlin; University of Saarland.

[Allen, 1983] Allen, J. (1983). Recognising intentions from natural language utterances. In *Computational Models of Discourse, Brady, M., and Berwick, R.C. (eds)*, pages 107–166. MIT Press, Cambridge, MA.

[Allen and Core, 1997] Allen, J. and Core, M. (1997). Draft of DAMSL: Dialog Act Markup in Several Layers. Available at `http://www.cs.rochester.edu/research/cisd/resources/damsl/`.

[Allen et al., 2001] Allen, J., Ferguson, G., and Stent, A. (2001). An architecture for more realistic conversational systems. In *Proceedings of the 6th International Conference on Intelligent User Interfaces (IUI'01), Santa Fe, New Mexico, USA*, pages 1–8.

[Allen and Perrault, 1980] Allen, J. and Perrault, C. (1980). Analyzing intention in dialogues. *Artificial Intelligence*, 15(3):143–178.

[Allen et al., 1994] Allen, J., Schubert, L., Fergusorr, G., Heeman, P., Hee Hwang, C., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., and Traum, D. (1994). The TRAINS project: a case study in building a conversational planning agent. TRAINS Technical Note 94-3, University of Rochester.

[Allwood, 1976] Allwood, J. (1976). Linguistic communication as action and cooperation. *Gothenburg Monographs in Linguistics 2, Göteborg University, Department of Linguistics*.

[Allwood, 1977] Allwood, J. (1977). A critical look at speech act theory. *Logic, Pragmatics and Grammar*, pages 53–69.

[Allwood, 1992] Allwood, J. (1992). On dialogue cohesion. *Gothenburg Papers in Theoretical Linguistics 65*.

[Allwood, 2000] Allwood, J. (2000). An activity-based approach to pragmatics. *Abduction, Belief and Context in Dialogue*, pages 47–81.

[Allwood, 2002] Allwood, J. (2002). Bodily communication - dimensions of expression and content. *Multimodality in Language and Speech Systems*, pages 7–26.

[Allwood et al., 2005] Allwood, J., Ahlsèn, E., Lund, J., and Sundqvist, J. (2005). Multimodality in Own Communication Management. In *Proceedings from the Second Nordic Conference on Multimodal Communication*, Göteborg.

[Allwood et al., 1997] Allwood, J., Ahlsèn, E., Nivre, J., and Larsson, S. (1997). *Own Communication Management: Kodningsmanual*. Göteborg University: Department of Linguistics.

[Allwood and Cerrato, 2003] Allwood, J. and Cerrato, L. (2003). A study of gestural feedback expressions. *Proceedings of the First Nordic Symposium on Multimodal Communication*, pages 7–22.

[Allwood et al., 1993] Allwood, J., Nivre, J., and E, A. (1993). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9-1:1–26.

[Allwood et al., 2000] Allwood, J., Traum, D., and Jokinen, K. (2000). Cooperation, dialogue and ethics. *International Journal of Human Computer Studies*, pages 871–914.

[AMI-Consortium, 2005a] AMI-Consortium (2005a). Coding guidelines for affect annotation of the AMI corpus. Available at `http://mmm.idiap.ch/private/ami/annotation/EmotionAnnotationManual-v1.0%.pdf`.

[AMI-Consortium, 2005b] AMI-Consortium (2005b). Guidelines for dialogue act and addressee annotation version 1.0. Available at `http://www.amiproject.org/`.

[Andry et al., 1990] Andry, F., Bilange, E., Charpentier, F., Choukri, K., Ponamalè, M., and Soudoplatoff, S. (1990). Computerised simulation tools for the design of an oral dialogue system. Selected Publications, 1988-1990, SUNDIAL Project (Esprit P2218). Commission of the European Communities.

[Ang et al., 2005] Ang, J., Liu, Y., and Shriberg, E. (2005). Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 1061–1064, Philadelphia, USA.

[Argyle, 1994] Argyle, M. (1994). *The psychology of interpersonal behaviour*. Penguin Books, London.

[Asher and Lascarides, 2003] Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

[Austin, 1962] Austin, J. (1962). *How to do things with words*. University Press, Oxford.

[Bach, 1981] Bach, E. (1981). On time, tense, and aspect: An essay in english metaphysics. In Cole, P., editor, *Radical Pragmatics*. Academic Press, New York.

[Barkhuysen et al., 2008] Barkhuysen, P., Krahmer, E., and Swerts, M. (2008). The interplay between auditory and visual cues for end-of-utterance detection. *The Journal of the Acoustical Society of America*, 123(1):354–365.

[Bavelas and Chovil, 2000] Bavelas, J. and Chovil, N. (2000). Visible acts of meaning. an integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology*, 19:163–194.

[Bilange, 1991] Bilange, E. (1991). A task independent oral dialogue model. In *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics*, pages 83–88, Berlin, Germany. Association for Computational Linguistics.

[Boersma and Weenink, 2009] Boersma, P. and Weenink, D. (2009). Praat: doing phonetics by computer. computer program. Available at `http://www.praat.org/`.

[Bos, 2002] Bos, J. (2002). *Underspecification and resolution in discourse semantics. PhD Thesis*. Saarland University, Saarbrücken.

[Bos et al., 2003] Bos, J., Klein, E., Lemon, O., and Oka, T. (2003). DIPPER: description and formalisation of an information-state update dialogue system architecture. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 115–124.

[Brown and Levinson, 1987] Brown, P. and Levinson, S. (1987). *Politeness - Some universals in language usage*. Cambridge University Press.

[Bunt, 1989] Bunt, H. (1989). Information dialogues as communicative action in relation to partner modelling and information processing. In Taylor, M., Neel, F., and Bouwhuis, D., editors, *The Structure of Multimodal Dialogue*, volume 1, pages 47–73. Elsevier, North Holland, The Netherlands.

[Bunt, 1992] Bunt, H. (1992). Belief contexts in human-computer dialogue. In Nauta, D., Nijholt, A., and Schaake, J., editors, *Pragmatics in Language Technology. Proceedings of 4th Twente Workshop on Language Technology*, pages 106 – 114, University of Twente, Enschede.

[Bunt, 1994] Bunt, H. (1994). Context and dialogue control. *THINK Quarterly 3(1)*, pages 19–31.

[Bunt, 1996] Bunt, H. (1996). Interaction management functions and context representation requirements. In Luperfoy, S., Nijholt, A., and Veldhuizen van Zanten, G., editors, *Dialogue Management in Natural Language Systems*, pages 187–198.

[Bunt, 1999] Bunt, H. (1999). Dynamic interpretation and dialogue theory. In Taylor, M., Neel, F., and D., B., editors, *The structure of multimodal dialogue II*, pages 139–166. John Benjamins, Amsterdam.

[Bunt, 2000] Bunt, H. (2000). Dialogue pragmatics and context specification. In Bunt, H. and Black, W., editors, *Abduction, Belief and Context in Dialogue; studies in computational pragmatics*, pages 81–105. John Benjamins, Amsterdam.

[Bunt, 2005] Bunt, H. (2005). A framework for dialogue act specification. In *Proceedings of the 1st Workshop on Interoperable Semantic Annotation*, Tilburg.

[Bunt, 2006] Bunt, H. (2006). Dimensions in dialogue act annotation. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 919–924, Genoa, Italy.

[Bunt, 2007a] Bunt, H. (2007a). Semantic underspecification: which techniques for what purpose? In Bunt, H. and Muskens, R., editors, *Computing Meaning, Vol. 3*, pages 55–85. Springer, Dordrecht.

[Bunt, 2007b] Bunt, H. (2007b). The semantics of semantic annotation. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation (PACLIC21)*, page 1328.

[Bunt, 2009a] Bunt, H. (2009a). The DIT++ taxonomy for functional dialogue markup. In Heylen, H., Pelachaud, C., Catizone, R., and Traum, D., editors, *Proceedings of the AAMAS 2009 Workshop 'Towards a Standard Markup Language for Embodied Dialogue Acts' (EDAML 2009)*, pages 13–25, Budapest.

[Bunt, 2009b] Bunt, H. (2009b). Multifunctionality and multidimensional dialogue semantics. In *Proceedings of the DiaHolmia Workshop on the Semantics and Pragmatics of Dialogue*, pages 3–15, Stockholm, Sweden.

[Bunt, 2009c] Bunt, H. (2009c). Semantic annotations as complementary to underspecified semantic representations. In *Proceedings of the 8th International Conference on Computational Semantics (IWCS-8)*, Tilburg, The Netherlands.

[Bunt, 2011] Bunt, H. (2011). The semantic of dialogue acts. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, Oxford, UK.

[Bunt and Girard., 2005] Bunt, H. and Girard., Y. (2005). Designing an open, multidimensional dialogue act taxonomy. In Gardent, C. and Gaiffe, B., editors, *Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue*, pages 37–44, Nancy.

[Bunt et al., 2007a] Bunt, H., Keizer, S., and Morante, R. (2007a). A computational model of grounding in dialogue. In *Proceedings of the Workshop in Discourse and Dialogue. Lecture Notes in Computer Science 4629*, pages 591–598, Antwerp, Belgium.

[Bunt and Overbeeke, 2008a] Bunt, H. and Overbeeke, C. (2008a). An extensible, compositional semantics of temporal annotation. In *Proceedings of LAW-II, the Second Linguistic Annotation Workshop*, Marrakech, Morocco.

[Bunt and Overbeeke, 2008b] Bunt, H. and Overbeeke, C. (2008b). Towards formal interpretation of semantic annotation. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

[Bunt et al., 2007b] Bunt, H., Petukhova, V., and Schiffrin, A. (2007b). LIRICS Deliverable D4.4. multilingual test suites for semantically annotated data. Available at `http://lirics.loria.fr`.

[Bunt and Romary, 2004] Bunt, H. and Romary, L. (2004). Standardization in multimodal content representation: Some methodological issues. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon.

[Bunt and Schiffrin, 2007] Bunt, H. and Schiffrin, A. (2007). Documented compilation of semantic data categories. deliverable d 4.3. Available at `http://lirics.loria.fr`.

[Campbell, 2008] Campbell, N. (2008). Individual traits of speaking style and speech rhythm in a spoken discourse. In *Verbal and non-verbal features of human-human and human-machine interaction. Revised papers from COST Action 2102 International Conference, Patras, Greece.* Berlin: Springer.

[Carberry, 1990] Carberry, S. (1990). *Plan recognition in natural language dialogue. ACL-MIT Press Series in Natural Language Processing*. Bradford Books, MIT Press, Cambridge, Massachusetts.

[Carletta, 1996] Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

[Carletta et al., 1996] Carletta, J. C., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., and Anderson, A. (1996). *HCRC Dialogue Structure Coding Manual*. Human Communication Research Centre HCRC TR-82, University of Edinburgh.

[Carlson et al., 2001] Carlson, L., Marcu, D., and Okurowski, M. (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10.

[Cassell et al., 2001] Cassell, J., Nakano, Y., and Bickmore, T. (2001). Non-verbal cues for discourse structure. In *Proceedings of Association for Computational Linguistics Annual Conference (ACL)*, pages 106–115.

[Cassell et al., 1999] Cassell, J., Torres, O., and Prevost, S. (1999). Turn taking vs. discourse structure: How best to model multimodal conversation. In Wilks, Y., editor, *Machine Conversations*, pages 143–154. Kluwer, The Hague.

[Clark, 1992] Clark, H. (1992). *Arenas of language use*. University of Chicago Press and CSLI.

[Clark, 1996] Clark, H. (1996). *Using language*. Cambridge University Press.

[Clark and Fox Tree, 2002] Clark, H. and Fox Tree, J. (2002). Using 'uh' and 'um' in spontaneous speech. *Cognition*, 84:73–111.

[Clark and Krych, 2004] Clark, H. and Krych, M. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50:62–81.

[Clark and Schaefer, 1989] Clark, H. and Schaefer, E. (1989). Contributing to discourse. *Cognitive Science*, 13:259–294.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20:37–46.

[Cohen and Levesque, 1990] Cohen, P. and Levesque, H. (1990). Persistence, intention and commitment. In *Intentions in Communication, Chapter 12*. Morgan Kaufmann.

[Cohen and Perrault, 1979] Cohen, P. and Perrault, C. (1979). Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212.

[Cohen, 1984] Cohen, R. (1984). A computational theory of the function of clue words in argument understanding. In *Proceedings of the Coling-ACL 1984*, pages 251–258, Standford.

[Cohen, 1995] Cohen, W. (1995). Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning (ICML'95)*, pages 115–123.

[Condon and Osgton, 1971] Condon, W. and Osgton, W. (1971). Speech and body motion synchrony of the speaker-hearer. In Horton, D. and Jenkins, J., editors, *The perception of language*, pages 150–184. Academic Press, New York.

[Cooper and Herskovits, 1992] Cooper, G. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

[Cooper, 2004] Cooper, R. (2004). A type-theoretic approach to information state update in issue-based dialogue management. In *Invited talk at CATALOG04, 8th Workshop on the Semantics and Pragmatics of Dialogue*, Barcelona, Spain.

[Corbett and Chang, 1983] Corbett, A. and Chang, F. (1983). Pronoun disambiguating: Accessing potential antecedents. *Memory and Cognition*, 11:283–294.

[Core and Allen, 1997] Core, M. and Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston.

[Craggs and McGee Wood, 2004] Craggs, R. and McGee Wood, M. (2004). A categorical annotation scheme for emotion in the linguistic content of dialogue. In Andre, E., Dybkjær, L., Minker, W., and Heisterkamp, P., editors, *Proceedings of the Affective Dialogue Systems, Tutorial and Research Workshop*, pages 89–100, Berlin. Springer.

[Cramer, 1985] Cramer, Y. (1985). Transcriptie dialoogexperiment. Report No 513, juni, Eindhoven: Institute for Perception Research.

[Daelemans et al., 1999] Daelemans, W., van den Bosch, A., and Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Machine Learning*, 34(1/3):11–43.

[Dahlbaeck and Jonsson, 1998] Dahlbaeck, N. and Jonsson, A. (1998). A coding manual for the linköping dialogue model. Unpublished manuscript.

[de Ruiter et al., 2006] de Ruiter, J., Mitterer, H., and Enfield, N. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82:515–535.

[DeVault and Stone, 2003] DeVault, D. and Stone, M. (2003). Domain inference in incremental interpretation. In *Proceedings of the Workshop on Inference in Computational Semantics*, pages 73–87, INRIA Lorraine, Nancy, France.

[Dhillon et al., 2004] Dhillon, R., Bhagat, S., Carvey, H., and Shriberg, E. (2004). Meeting recorder project: dialogue labelling guide. ICSI Technical Report TR-04-002.

[Di Eugenio et al., 1998] Di Eugenio, B., Jordan, P., and Pylkkaenen, L. (1998). The COCONUT project: dialogue annotation manual. ISP Technical Report 98-1.

[Dietterich, 2002] Dietterich, T. (2002). Machine learning for sequential data: a review. In Caelli, T., Amin, A., Duin, R., Kamel, M., and Ridder, D., editors, *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30.

[Duncan, 1970] Duncan, S. (1970). Towards a grammar for floor apportionment: A system approach to face-to-face interaction. In *Proceedings of the Second Annual Environmental Design Research Association Conference*, pages 225–236, Philadelphia.

[Duncan and Fiske, 1977] Duncan, S. and Fiske, D. (1977). *Face-to-face interaction: research, methods, and theory*. Lawrence Erlbaum Associates, New Jersey.

[Ekman, 1972] Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In Cole, J., editor, *Nebraska Symposium on Motivation*, pages 207–283. University of Nebraska Press, Lincoln, Nebraska.

[Ekman, 1999] Ekman, P. (1999). Basic emotions. In Dalgleish, T. and Power, M., editors, *Handbook of Cognition and Emotion*. John Wiley and Sons, Ltd., Sussex, U.K.

[Ekman and Friesen, 1969] Ekman, P. and Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1:49–98.

[Erickson, 1975] Erickson, F. (1975). One function of proxemic shifts in face-to-face interaction. In Kendon, A., Harris, R., and Key, M., editors, *Organization of behavior in face-to-face interaction*, pages 175–187. Mouton, Den Haag.

[Exline, 1963] Exline, R. (1963). Exploration in the process of person perception: visual interaction in relation to competitions, sex and need for affiliation. *Journal of Personality*, 31 (1):1–20.

[Fagin et al., 1995] Fagin, R., Halpern, J., Moses, Y., and Vardi, M. (1995). *Reasoning About Knowledge*. MIT Press.

[Fernandez and Picard, 2002] Fernandez, R. and Picard, R. W. (2002). Dialog act classification from prosodic features using Support Vector Machines. In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France.

[Ford and Thompson, 1996] Ford, C. and Thompson, S. (1996). Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In Schegloff, E. and Thompson, S., editors, *Interaction and grammar*, pages 135–184. Cambridge: Cambridge University Press.

[Geertzen, 2007] Geertzen, J. (2007). Ditat: a flexible tool to support web-based dialogue annotation. In Geertzen, J., Thijsse, E., Bunt, H., and Schiffrin, A., editors, *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7)*, pages 320–323, Tilburg University, Tilburg, The Netherlands.

[Geertzen, 2009] Geertzen, J. (2009). *Dialogue act recognition and prediction: exploration in computational dialogue modelling. PhD Thesis*. Tilburg University, The Netherlands.

[Geertzen and Bunt, 2006] Geertzen, J. and Bunt, H. (2006). Measuring annotator agreement in a complex hierarchical dialogue act scheme. In *Proceedings of the 7th SigDial Workshop on Discourse and Dialogue, Sidney*, pages 126–133. Association for Computational Linguistics.

[Geertzen et al., 2004] Geertzen, J., Girard, Y., and Morante, R. (2004). The DIAMOND project. Poster at the 8th Workshop on the Semantics and Pragmatics of Dialogue (CATALOG 2004), Barcelona, Spain.

[Geertzen et al., 2007] Geertzen, J., Petukhova, V., and Bunt, H. (2007). A multidimensional approach to utterance segmentation and dialogue act classification. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 140–149, Antwerp, Belgium. Association for Computational Linguistics.

[Geertzen et al., 2008] Geertzen, J., Petukhova, V., and Bunt, H. (2008). Evaluating dialogue act tagging with naive and expert annotators. In (ELRA), E. L. R. A., editor, *Proceedings of the 6th International Language Resources and Evaluation (LREC 2008)*, pages 1076–1082, Marrakech, Morocco.

[Ginzburg, 1998] Ginzburg, J. (1998). Clarifying utterances. In *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues*, pages 11–30, Enschede, The Netherlands.

[Goffman, 1963] Goffman, E. (1963). *Behavior in Public Places*. Basic Books, New York.

[Goodwin, 1981] Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. Academic Press, New York.

[Gravano et al., 2007] Gravano, A., Benus, S., Chavez, H., Hirschberg, J., and Wilcox, L. (2007). On the role of context and prosody in the interpretation of 'okay'. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 800–807, Prague, Czech Republic.

[Grosjean and Hirt, 1996] Grosjean, F. and Hirt, C. (1996). Using prosody to predict the end of sentences in english and french: Normal and brain-damaged subjects. *Language and Cognitive Processes*, 11:107–134.

[Grosz and Sidner, 1986] Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204.

[Grosz and Sidner, 1990] Grosz, B. J. and Sidner, C. L. (1990). Plans for discourse. In Cohen, P., Morgan, J., and Pollack, M., editors, *Intentions in Communication*, pages 417–444. MIT Press, Cambridge, Massachusetts.

[Gupta et al., 2007] Gupta, S., Niekrasz, J., Purver, M., and Jurafsky, D. (2007). Resolving 'you' in multi-party dialog. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 227–230, Antwerp, Belgium.

[Gut et al., 2003] Gut, U., Looks, K., Thies, A., and Gibbon, D. (2003). CoGesT: Conversational gesture transcription system. Version 1.0. Technical report. Bielefeld University.

[Hadar et al., 1984] Hadar, U., Steiner, T., Grant, E., and Rose, F. (1984). The timing of shifts of head postures during conversations. *Human Movement Science*, 3:237–245.

[Haddock, 1989] Haddock, N. (1989). Computational models of incremental semantic interpretation. *Language and Cognitive Processes*, 14 (3):SI337–SI380.

[Hartmann et al., 2005] Hartmann, B., Mancini, M., Buisine, S., and Pelachaud, C. (2005). Design and evaluation of expressive gesture synthesis for embodied conversational agents. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1095–1096, Utretch, The Netherlands.

[Heeman and Allen, 1999] Heeman, P. and Allen, J. (1999). Speech repairs, intonational phrases and discourse markers: Modelling speakers utterances in spoken dialogue. *Computational Linguistics*, 12(3):1–45.

[Heylen, 2006] Heylen, D. (2006). Head gestures, gaze and the principles of conversational structure. *International journal of Humanoid Robotics*, 3(3):241–267. ISSN=0219-8436.

[Hintikka, 1962] Hintikka, J. (1962). *Knowledge and Belief*. Cornell University Press, Ithaca, NY.

[Hirsch, 1989] Hirsch, R. (1989). *Argumentation, information, and interaction: Studies in face-to-face interactive argumentation under differing turn-taking conditions*. Department of Linguistics, University of Göteborg.

[Hirschberg and Litman, 1993] Hirschberg, J. and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 25(4):501–530.

[Hobbs, 1979] Hobbs, J. (1979). Coherence and coreference. *Cognitive Science*, 3:67–90.

[Hobbs, 1985a] Hobbs, J. (1985a). On the coherence and structure of discourse. Research Report 85-37, CSLI, Stanford.

[Hobbs, 1985b] Hobbs, J. (1985b). Ontological promiscuity. In *Proceedings 23rd Annual Meeting of the ACL*, pages 61–69, Chicago.

[Hockey, 1993] Hockey, B. (1993). Prosody and the role of okay and uh-huh in discourse. In *Proceedings of the Eastern States Conference on Linguistics*, pages 128–136.

[Hovy, 1990] Hovy, E. (1990). Approaches to the planning of coherent text. In Swartout, C. and Mann, W., editors, *Natural Language in Artificial Intelligence and Computational Linguistics*, pages 83–102. Kluwer, Boston.

[Hovy, 1995] Hovy, E. (1995). The multifunctionality of discourse markers. In *Proceedings of the Workshop on Discourse Markers*, Egmond-aan-Zee, The Netherlands.

[Hovy and Maier, 1995] Hovy, E. and Maier, E. (1995). Parsimonious of profligate: how many and which discourse structure relations? unpublished manuscript.

[Ichikawa, 1998] Ichikawa, A. e. a. (1998). Standardising annotation schemes for japanese discourse. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 731–736, Spain.

[ISO, 2009] ISO (2009). *ISO 24612:2009 Language resource management: Linguistic annotation framework (LAF)*. ISO, Geneva.

[ISO, 2010] ISO (2010). *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO DIS 24617-2*. ISO Central Secretariat, Geneva.

[Jurafsky et al., 1998a] Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., and van Ess-Dykema, C. (1998a). Automatic detection of discourse structure for speech recognition and understanding. In *Proceedings of the IEEE Workshop on Speech Recognition and Understanding*, pages 88–95, Santa Barbara, USA.

[Jurafsky et al., 1997] Jurafsky, D., Shriberg, E., and Biasca, D. (1997). *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation: Coders Manual, Draft 13*. University of Colorado.

[Jurafsky et al., 1998b] Jurafsky, D., Shriberg, E., Fox, B., and Curl, T. (1998b). Lexical, prosodic, and syntactic cues for dialogue acts. In Stede, M., Wanner, L., and Hovy, E., editors, *Discourse Relations and Discourse Markers: Proceedings of the Conference*, pages 114–120, Somerset, New Jersey, USA. Association for Computational Linguistics.

[Keizer, 2003] Keizer, S. (2003). *Reasoning under uncertainty in natural language dialogue using Bayesian Networks. PhD Thesis*. Twente University Press, The Netherlands.

[Keizer and Bunt, 2006] Keizer, S. and Bunt, H. (2006). Multidimensional dialogue management. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 37–45, Sydney.

[Keizer and Bunt, 2007] Keizer, S. and Bunt, H. (2007). Evaluating combinations of dialogue acts for generation. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 158–165, Antwerp.

[Keizer et al., 2011] Keizer, S., Bunt, H., and Petukhova, V. (2011). Multidimensional dialogue management. In van den Bosch, A. and Bouma, G., editors, *IMIX book*. Springer.

[Keizer and Morante, 2007] Keizer, S. and Morante, R. (2007). Dialogue simulation and context dynamics for dialogue management. In *Proceedings of the 16th Nordic Conference of Computational Linguistics, NODALIDA 2007*, pages 310–317, University of Tartu, Estonia.

[Kendon, 1967] Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.

[Kendon, 2004] Kendon, A. (2004). *Gesture: visible action as utterance.* Cambridge University Press, Cambridge.

[Klein, 1999] Klein, M. (1999). Standardization efforts on the level of dialogue act in the MATE project. Available at `http://acl.ldc.upenn.edu/W/W99/W99-0305.pdf`.

[Klein and Soria, 1998] Klein, M. and Soria, C. (1998). Supported coding schemes. MATE Deliverable D1.1.

[Konolidge, 1986] Konolidge, K. (1986). *A Deduction Model of Belief*. Pitman Publishing.

[Krippendorff, 1980] Krippendorff, K. (1980). *Content analysis: an introduction to its methodology*, volume 5. Sage Publications, Beverly Hills, London.

[Lakoff, 1973] Lakoff, R. (1973). The logic of politeness: or minding your P's and Q's. In Corum, C., Smith-Stark, T., and A., W., editors, *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society*, pages 292–305. Chicago Linguistic Society.

[Larsson, 1997] Larsson, S. (1997). A type hierarchy of dialogue moves. Available at `http://www.ling.gu.se/~sl/sdime/sdime_type.html`.

[Larsson, 1998] Larsson, S. (1998). Coding schemas for dialogue moves. technical report from the S-DIME project. Available at `http://www.ling.gu.se/~sl/papers.html`.

[Larsson, 2002] Larsson, S. (2002). *Issue-based dialogue management. PhD thesis*. Göteborg University, Göteborg, Sweden.

[Larsson et al., 2000] Larsson, S., Ljunglöf, P., Cooper, R., Engdahl, E., and Ericsson, S. (2000). GoDis: an accommodating dialogue system. In *Proceedings ANLP/NAACL 2000 Workshop on Conversational Systems, Seattle, Washington*, pages 7–10.

[Larsson and Traum, 2000] Larsson, S. and Traum, D. (2000). Information state and dialogue management in the Trindi dialogue move engine toolkit. *Natural Language Engineering*, 6(3-4):323–340.

[Laskowski and Burger, 2005] Laskowski, K. and Burger, S. (2005). Annotation and analysis of emotionally relevant behavior in the ISL meeting corpus. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1111–1116, Genoa, Italy.

[Lee et al., 2004] Lee, K., Burnard, L., Romary, L., Bunt, H., Declerck, T., Erjavec, T., and Clergerie, d. l. E. (2004). Towards an international standard in feature structure representation. In Declerck, T., Ide, N., and Romary, L., editors, *Proceedings of the Workshop 'A Registry of Linguistic Data Categories within an Integrated Language Resources Repository'*, pages 63–70, Lisbon.

[Leech, 1971] Leech, G. (1971). *Meaning and the English verb*. Longman, London.

[Lendvai et al., 2003] Lendvai, P., Bosch, v. d. A., and Krahmer, E. (2003). Machine learning for shallow interpretation of user utterances in spoken dialogue systems. In *Proceedings of the EACL-03 Workshop on Dialogue Systems: interaction, adaptation and styles of management*, pages 69–78, Budapest.

[Lendvai et al., 2004] Lendvai, P., Bosch, v. d. A., Krahmer, E., and Canisius, S. (2004). Memory-based robust interpretation of recognised speech. In *Proceedings of the 9th International Conference on Speech and Computer (SPECOM '04)*, pages 415–422, St. Petersburgh, Russia.

[Lendvai and Geertzen, 2007] Lendvai, P. and Geertzen, J. (2007). Token-based chunking of turn-internal dialogue act sequences. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 174–181, Antwerp, Belgium.

[Lesch et al., 2005] Lesch, S., Kleinbauer, T., and Alexandersson, J. (2005). A new metric for the evaluation of dialog act classification. In *Proceedings of 9th Workshop on the semantics and pragmatics of dialogue (Dialor)*, Nancy, France.

[Levelt, 1989] Levelt, W. (1989). *Speaking: from intention to articulation*. MIT Press, Cambridge, Massachusetts, etc.

[Levinson, 1983] Levinson, S. (1983). *Pragmatics*. Cambridge University Press, Cambridge, UK.

[Lewin, 1998] Lewin, I. (1998). The autoroute dialogue demonstrator. Technical Report CRC-073, SRI Cambridge Computer Science Research Centre.

[Louwerse and Mitchell, 2003] Louwerse, M. and Mitchell, H. (2003). Toward a taxonomy of a set of discourse markers in dialogue: A theoretical and computational linguistic account. *Discourse Processes*, 35(3):243–281.

[Mann and Thompson, 1988] Mann, W. and Thompson, S. (1988). *Rhetorical structure theory: toward a functional theory of text organisation*. The MIT Press, Cambridge, MA.

[Mazeland, 2003] Mazeland, H. (2003). *Inleiding in de conversatie-analyse*. Uitgeverij Coutinho, Bussum, The Netherlands.

[McClave, 2001] McClave, E. (2001). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878.

[McNeill, 1992] McNeill, D. (1992). *Hand and Mind: what gestures reveal about thought*. The University of Chicago Press, Chicago and London.

[Miltsakaki et al., 2004] Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004). The Penn Discourse Treebank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon.

[Milward and Cooper, 2009] Milward, D. and Cooper, R. (2009). Incremental interpretation: applications, theory, and relationship to dynamic semantics. In *Proceedings COLING 2009, Kyoto, Japan*, pages 748–754.

[Mindt, 1998] Mindt, D. (1998). *An empirical grammar of the English verb: modal verbs*. Cornelson, Berlin.

[Moore and Pollack, 1992] Moore, J. and Pollack, M. (1992). A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18:537–544.

[Moore, 1985] Moore, R. (1985). A formal theory of knowledge and action. In *Formal Theories of the Commonsense World*, pages 319–358. Ablex Publishing, Norwood, NJ.

[Morante, 2007] Morante, R. (2007). *Computing meaning in interaction. PhD Thesis*. Tilburg University, The Netherlands.

[Muller and Prévot, 2003] Muller, P. and Prévot, L. (2003). An empirical study of acknowledgment structures. In *Proceedings of Diabruck, 7th Workshop on Semantics and Pragmatics of Dialogue, Saarbrücken*.

[Nagata and Morimoto, 1994] Nagata, M. and Morimoto, T. (1994). First steps toward statistical modelling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15:193–203.

[Nakano et al., 1999] Nakano, M., Miyazaki, N., Hirasawa, J., Dohsaka, K., and Kawabata, T. (1999). Understanding unsegmented user utterances in real-time spoken dialogue systems. In *Proceedings of the 37th Annual Conference of the Association of Computational Linguistics, ACL*, pages 200–207.

[Nakano et al., 2003] Nakano, Y., Reinstein, G., Stocky, T., and Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 553–561.

[Nakatani and Traum, 1999] Nakatani, C. and Traum, D. (1999). *Draft: discourse structure coding manual. Version 1.0. Technical Report UMIACS-TR-99-03*. University of Maryland.

[Nivre et al., 1998] Nivre, J., Allwood, J., and Ahlsèn, E. (1998). *Interactive Communication Management: Coding Manual. Version 1.0.* Göteborg University: Department of Linguistics.

[Nöth et al., 2002] Nöth, N., Batliner, A., Warnke, V., Haas, J.-P., Boros, M., Buckow, J., Huber, R., Gallwitz, F., Nutt, M., and Niemann, H. (2002). On the use of prosody in automatic dialogue understanding. *Speech Communication*, 36(1-2):45–62.

[Novick et al., 1996] Novick, D., Hansen, B., and Ward, K. (1996). Coordinating turn-taking with gaze. In *Proceedings of the International Symposium on Spoken Dialogue*, pages 53–56, Philadelphia, PA.

[Ordelman and Heylen, 2005] Ordelman, R. and Heylen, D. (2005). Annotation of emotions in meetings in the AMI project. In *Presentation at the HUMAINE WP6 Workshop*.

[Padilha and Carletta, 2003] Padilha, E. and Carletta, J. (2003). Nonverbal behaviours improving a simulation of small group discussion. In *Proceedings of the First International Nordic Symposium of Multimodal Communication*, pages 93–105, Copenhagen, Denmark.

[Pavelin, 2002] Pavelin, B. (2002). *Le Geste à la parole*. Presses Universitaires du Mirail, Toulouse.

[Petukhova, 2005] Petukhova, V. (2005). *Multidimensional interaction of multimodal dialogue acts in meetings. MA thesis,*. Tilburg University.

[Petukhova and Bunt, 2007] Petukhova, V. and Bunt, H. (2007). A multidimensional approach to multimodal dialogue act annotation. In *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS)*, pages 142–153.

[Petukhova and Bunt., 2009] Petukhova, V. and Bunt., H. (2009). Dimensions in communication. TiCC Technical Report TR 2009-002, Tilburg University.

[Petukhova and Bunt, 2009a] Petukhova, V. and Bunt, H. (2009a). Grounding by nodding. In *Proceedings of the 1st Conference on Gesture and Speech in Interaction*, Poznan, Poland.

[Petukhova and Bunt, 2009b] Petukhova, V. and Bunt, H. (2009b). Towards a multidimensional semantics of discourse markers in spoken dialogue. In Bunt, H., Petukhova, V., and S., W., editors, *Proceedings of the Eighth International Conference on Computational Semantics (IWCS)*, pages 157–168, Tilburg.

[Petukhova and Bunt, 2009c] Petukhova, V. and Bunt, H. (2009c). Who's next? speaker-selection mechanisms in multiparty dialogue. In *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue*, pages 19–26, Stockholm,.

[Petukhova and Bunt, 2010a] Petukhova, V. and Bunt, H. (2010a). Context-driven dialogue act generation. In Łupkowski, P. and Purver, M., editors, *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue: Aspects of semantics and pragmatics of dialogue*, pages 151–153, Poznan, Poland.

[Petukhova and Bunt, 2010b] Petukhova, V. and Bunt, H. (2010b). Introducing communicative function qualifiers. In Fang A., I. N. and J., W., editors, *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, pages 123–133, Hong Kong.

[Petukhova and Bunt, 2010c] Petukhova, V. and Bunt, H. (2010c). Towards an integrated scheme for semantic annotation of multimodal dialogue data. In *Proceedings of the seventh international conference on language resources and evaluation*, pages 2556–2563. Paris: ELRA.

[Petukhova and Bunt, 2011] Petukhova, V. and Bunt, H. (2011). Incremental dialogue act understanding. In Bos, J. and Pulman, S., editors, *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, pages 235–245, Oxford, UK.

[Petukhova et al., 2010] Petukhova, V., Bunt, H., and Malchanau, A. (2010). Empirical and theoretical constraints on dialogue act combinations. In Łupkowski, P. and Purver, M., editors, *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue: Aspects of semantics and pragmatics of dialogue*, pages 1–8, Poznan, Poland.

[Petukhova et al., 2011] Petukhova, V., Prévot, L., and Bunt, H. (2011). Multi-level discourse relations between dialogue units. In *Proceedings of the Sixth Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 18–28, Oxford: Oxford University.

[Pinkal, 1999] Pinkal, M. (1999). On semantic underspecification. In Bunt, H. and Muskens, R., editors, *Computing Meaning*, volume 1, pages 33–56. Kluwer, Dordrecht.

[Poesio and Traum, 1998] Poesio, M. and Traum, D. (1998). Towards an axiomatization of dialogue acts. In *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues*, pages 207–222.

[Polanyi, 1988] Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638.

[Popescu-Belis, 2004] Popescu-Belis, A. (2004). Dialogue act tagsets for meeting understanding: an abstraction based on the DAMSL, Switchboard and ICSI-MR tagsets. Technical report, IM2.MDM-09, v1.2.

[Popescu-Belis, 2005] Popescu-Belis, A. (2005). Dialogue acts: One or more dimensions? *IOSSCO Working Paper 62*.

[Popescu-Belis and Zufferey, 2006] Popescu-Belis, A. and Zufferey, S. (2006). Automatic identification of discourse markers in multiparty dialogues. Working paper 65. ISSCO, University of Geneva.

[Potts, 2003] Potts, C. (2003). *The Logic of Conventional Implicatures, PhD thesis*. UC Santa Cruz.

[Prasad et al., 2008] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc.

[Prüst et al., 1984] Prüst, H., Minnen, G., and Beun, R. (1984). Transcriptie dialoogexperiment. Report No 481, juni/juli, Eindhoven: Institute for Perception Research.

[Quek et al., 2000] Quek, F., McNeill, D., and Bryll, R. (2000). Gesture, speech, and gaze cues for discourse segmentation. In *Proceedings of the Computer Vision and Pattern Recognition CVPR*, volume 2, pages 247–254.

[Reese et al., 2007] Reese, B., Denis, P., Asher, N., Baldridge, J., and Hunter, J. (2007). Reference manual for the analysis and annotation of rhetorical structure. Version 1.0, University of Texas.

[Reidsma, 2008] Reidsma, D. (2008). *Annotations and subjective machines. Of annotators, embodied agents, users, and other humans. PhD Thesis*. University of Twente, Enschede, The Netherlands.

[Reidsma et al., 2006] Reidsma, D., Heylen, D., and Ordelman, R. (2006). Annotating emotion in meetings. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1117–1122, Genoa, Italy.

[Reithinger, 1997] Reithinger, N.and Klesen, M. (1997). Dialogue act classification using language models. In *Proceedings of EuroSpeech-97*, pages 2235–2238.

[Rienks and Verbree, 2005] Rienks, R. and Verbree, D. (2005). *Twente Argument Schema Annotation Manual v 0.99b*. University of Twente.

[Roque et al., 2006] Roque, A., Leuski, A., Rangarajan, V., Robinson, S., Vaswani, S., Narayanan, S., and Traum, D. (2006). Radiobot-CFF: A spoken dialogue system for military training. In *Proceedings of the Interspeech, 2006*, Pittsburgh, PA.

[Sacks et al., 1974] Sacks, H., Schegloff, E., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

[Sadek, 1991] Sadek, D. (1991). Dialogue acts are rational plans. In *Proceedings of the ESCA/ETRW Workshop on the Structure of Multimodal Dialogue*, pages 19–48, Maratea, Italy.

[Samuel et al., 1998] Samuel, K., Carberry, S., and K., V.-S. (1998). Dialogue act tagging with transformation-based learning. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 2, pages 1150 – 1156, Montreal.

[Sanders et al., 1992] Sanders, T., Spooren, W., and Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15:1–35.

[Scheflen, 1964] Scheflen, A. (1964). The significance of posture in communication systems. *Psychiatry*, 17:316–331.

[Schegloff, 1968] Schegloff, E. (1968). Sequencing in conversational openings. *American Anthropologist*, 70:1075–1095.

[Schiffrin, 1987] Schiffrin, D. (1987). *Discourse Markers*. Cambridge University Press, Cambridge.

[Searle, 1969] Searle, J. (1969). *Speech acts*. Cambridge University Press, Cambridge.

[Sedivy, 2003] Sedivy, J. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32(1):3–23.

[Sedivy et al., 1999] Sedivy, J., Tanenhaus, M., Chambers, C., and Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71:109–147.

[Selting, 2000] Selting, M. (2000). The construction of units in conversational talk. *Language in Society*, 29:477–517.

[Shriberg et al., 1998] Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., and Ess-Dykema van, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech (Special Issue on Prosody and Conversation)*, 41(3-4):439–487.

[Sidner and Israel, 1981] Sidner, C. and Israel, D. (1981). Recognizing intended meaning and speaker's plans. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 203–208, Vancouver, Canada.

[Simpson, 1994] Simpson, G. (1994). Context and the processing of ambiguous words. In Gernsbacher, M., editor, *Handbook of Psycholinguistics*, pages 359–374. Academic Press.

[Soria and Pirrelli, 2003] Soria, C. and Pirrelli, V. (2003). A multi-level annotation metascheme for dialogue acts. In Zampoli, A., Calzolari, N., and L., C., editors, *Computational Linguistics in Pisa. Linguistica Computazionale, Special Issue*, volume XVIII-XIX, pages 865–900. Pisa-Roma: IEPI.

[Stede, 2008] Stede, M. (2008). Disambiguating rhetorical structure. *Research on Language and Computation*, 6(3-4):311–332.

[Stent, 2000] Stent, A. (2000). The Monroe corpus. Technical Report TR728/TN99-2.

[Stolcke et al., 2000] Stolcke, A., Ries, K., Coccaro, K., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema van, C., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

[Swanson, 2010] Swanson, E. (2010). How not to theorize about the language of subjective uncertainty. In *Andy Egan and Brian Weatherson (eds.), Epistemic Modality*. Oxford University Press.

[Swerts, 1998] Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30:485–496.

[Swerts and Ostendorf, 1997] Swerts, M. and Ostendorf, M. (1997). Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Communication*, 22:25–41.

[Swinney, 1979] Swinney, D. (1979). Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behaviour*, 18:545–567.

[Tanenhaus et al., 1995] Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., and Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

[ten Bosch et al., 2004] ten Bosch, L., Oostdijk, N., and de Ruiter, J. (2004). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In *Lecture Notes in Artificial Intelligence LNCS/LNAI 3206, Brno, Czech Republic*, pages 563–570. Spinger Verlag.

[Tomita, 1986] Tomita, M. (1986). *Efficient parsing for natural language*. Kluwer, Dordrecht.

[Traum, 1994] Traum, D. (1994). A computational theory of grounding in natural language conversation. PhD Thesis. Dep. of Computer Science, University of Rochester.

[Traum, 1999] Traum, D. (1999). Computational models of grounding in collaborative systems. In Brennen, S.E. Giboin, A. and Traum, D., editors, *Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, pages 124–131, Menlo Park, CA, USA. American Association for Artificial Intelligence.

[Traum, 2000] Traum, D. (2000). 20 questions on dialogue act taxonomies. *Journal of Semantics*, 17(1):7–30.

[Traum and Allen, 1992] Traum, D. and Allen, J. (1992). A speech acts approach to grounding in conversation. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92)*, page 13740.

[Traum et al., 1999] Traum, D., Bos, J., Cooper, R., Larsson, S., Lewin, I., Matheson, C., and Poesio, M. (1999). A model of dialogue moves and information state revision. TRINDI project deliverable D2.1.

[Traum and Heeman, 1997] Traum, D. and Heeman, P. (1997). Utterance units in spoken dialogue. In *Proceedings of ECAI Workshop on Dialogue Processing in Social Language Systems*, pages 125–140, London, UK.

[Van den Bosch, 1997] Van den Bosch, A. (1997). *Learning to pronounce written words: A study in inductive language learning. PhD thesis*. Maastricht University, The Netherlands.

[van Dijk, 1979] van Dijk, T. A. (1979). Pragmatic connectives. *Journal of Pragmatics*, 3:447–456.

[Vark van et al., 1996] Vark van, R., Vreught de, J., and Rothkrantz, L. (1996). Analysing OVR dialogue coding scheme 1.0. Report 96-137.

[Vertegaal et al., 2001] Vertegaal, R., Slagter, R., Van der Veer, G., and Nijholt, A. (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proceedings of CHI'01: ACM Conference on Human Factors in Computing Systems*, pages 301–307, Seattle, WA.

[von Fintel and Gillies, 2007] von Fintel, K. and Gillies, A. (2007). An opinionated guide to epistemic modality. In Gendler, T. and Hawthorne, J., editors, *Oxford Studies in Epistemology*, volume 2, pages 32–62. Oxford University Press.

[Wahlster, 2000] Wahlster, W., editor (2000). *Verbmobil: foundation of speech-to-speech translation*. Springer, Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo.

[Webb et al., 2005] Webb, N., Hepple, M., and Wilks, Y. (2005). Error analysis of dialogue act classification. In *Proceedings of the 8th International Conference on Text, Speech and Dialogue*, volume 3658, pages 451–458, Karlovy Vary, Czech Republic.

[Wiemann and M.L., 1975] Wiemann, J. and M.L., K. (1975). Turn-taking in conversations. *Journal of Communication*, 25(2):75–92.

[Wilks and Ballim, 1991] Wilks, Y. and Ballim, A. (1991). Beliefs, stereotypes and dynamic agent modeling. In *User Modeling and User-Adapted Interaction*, volume 1(1), pages 33–65. Kluwer Academic Publishers, Dordrecht.

[Witten and Frank, 2000] Witten, I. and Frank, E. (2000). *Data Mining*. Morgan Kaufmann Publishers, San Francisco.

[Włodarczak, 2009] Włodarczak, M. (2009). Ranked multidimensional dialogue act annotation. MA thesis, Adam Mickiewicz University, Poznan.

[Włodarczak et al., 2010] Włodarczak, M., Bunt, H., and Petukhova, V. (2010). Entailed feedback: evidence from a ranking experiment. In Łupkowski, P. and Purver, M., editors, *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue: Aspects of semantics and pragmatics of dialogue*, pages 159–162, Poznan, Poland.

[Wolf and Gibson, 2005] Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

[Woszczyna and Waibel, 1994] Woszczyna, M. and Waibel, A. (1994). Inferring linguistic structure in spoken language. In *Proceedings of ICSLP-94*, pages 847–850, Yokohama, Japan.

[Xu et al., 2005] Xu, W., Carletta, J., Kilgour, J., and Karaiskos, V. (2005). *Coding Instructions for Topic Segmentation of the AMI Meeting Corpus (version 1.1)*. School of Informatics, University of Edinburgh.

[Zimmermann et al., 2005] Zimmermann, M., Lui, Y., Shriberg, E., and Stolcke, A. (2005). Toward joint segmentation and classification of dialog acts in multiparty meetings. In *Proceedings of the Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI-05)*, pages 187–193. Springer.

Summary

The main goal of this thesis is to contribute to the development of well-founded computational models of dialogue. We investigate natural spoken and multimodal dialogue behaviour both analytically and empirically. The approach that we present combines a multidimensional view on communication with a structured representation of dialogue context.

The main contributions of this thesis are: (1) the definition of a theoretically and empirically well-founded notion of dimension in dialogue act analysis, which provides a basis for the choice of dimensions in multidimensional dialogue act taxonomies and annotation schemes; (2) an empirically-based analysis of the multifunctionality of dialogue utterances, as observed in corpus data; (3) the identification and successful application of features of human non-verbal behaviour in the study of certain classes of dialogue acts, such as feedback acts, turn management acts, and discourse structuring acts; (4) the development of a machine learning-based approach to the incremental understanding of dialogue utterances, with a focus on the recognition of their communicative functions; (5) a context-driven approach to dialogue act interpretation and generation which enables the construction of intentionally multifunctional dialogue contributions.

We show in this thesis that the systematic application of a multidimensional view on communication leads to a better understanding of human dialogue behaviour and enables better computational modelling of multimodal dialogue. A range of problems can be solved when using a multidimensional approach to dialogue analysis and modelling, based on well-defined concepts of 'dimension' and 'dialogue act', supported by detailed empirical analysis of dialogue behaviour. The obtained insights in the nature, forms and linguistic and non-linguistic manifestation of multifunctionality opens the perspective of a dialogue system that understands and generates utterances which are multifunctional by design.

Chapter 3 defines the notion of 'dimension' that has a conceptual, theoretical and empirical significance not only for dialogue act annotation, but also for dialogue segmentation and interpretation, and that enables a more adequate dialogue modelling, since dimensions carry an essential part of the meaning of many dialogue utterances. We formulate general criteria to (1) decide on the kind of elements that should be included in the reference set of dimensions and why; and (2) how they can be organized in a taxonomy. We show how these criteria can be turned into operational tests for effectively making well-founded decisions for the design of a well-founded set of dimensions. Dimensions are then considered which correspond to well-studied communicative activities that dialogue participants perform and are distinguishable according to empirically observable behaviour in dialogue. Application of the criteria has led to the foundations of the set of ten dimensions in the DIT^{++} dialogue acts taxonomy and to the choice of nine dimensions for the ISO 24617-2 dialogue act annotation scheme.

In Chapter 4 we address the dialogue act annotation task. Multi- and one-dimensional approaches to this task are discussed and compared. From this discussion it has been concluded that multidimensional dialogue act annotation schemes do not only better capture fine-grained theoretical and empirical distinction of defined concepts resulting in better coverage of dialogue phenomena, are flexible and easy to adapt to various purposes and tasks domain, but also, contrary to what was generally believed, can be reliably applied by annotators. The multidimensional approach applied to dialogue act segmentation solves various notorious problems caused by disfluent speech, overlapping and simultaneous talk, and discontinuity of the segments that are relevant for analysis. We also show that a multidimensional approach to segmentation results in a more accurate analysis expressed in higher scores for automatic dialogue

act classification (Chapter 7).

We propose improvements and extensions of existing dialogue act annotation schemes. The first extension is concerned with relations that dialogue unit of various kinds may have. In studying the occurrence of discourse relations in dialogue, we have observed at least four types of relations: rhetorical relations between dialogue acts or between their semantic contents (interpropositional rhetorical relations); feedback dependence relations; and functional dependence relations between dialogue acts. Some of these relations may also involve larger units or groups of those, and we establish that the various kinds of relation show significant differences in scope and distance of attachment. A metamodel for dialogue act annotation is designed as an extension of the ISO 24617-2 metamodel, containing the various kinds of units in dialogue and the possible relations between them. Another extension concerns the representation of dialogue acts which involve uncertainty, conditionality or sentiment. We propose a set of qualifiers that can be attached to a communicative function in order to describe the speaker's behaviour more accurately, taking these aspects into account.

In Chapter 5 we investigate the forms of multifunctionality that occur in natural dialogue. The various functions that an utterance may have re often related to different communicative aspects ('dimensions'). The relation between possible forms of multifunctionality and conceptually distinguishable dimensions of communication are studied in detail. We do this analytically by studying possible semantic relations between communicative functions, both logical and pragmatic; and empirically, analysing the mutltifunctionality that actually occur in various types of dialogue units, such as single functional segments, embeded segments, and segment sequences. The results of this particular study do not only have consequences for the semantic interpretation of dialogue contributions, but also for their generation by spoken dialogue systems.

Chapter 6 is concerned with the interpretation of communicative behaviour that it is observed in the annotated dialogue corpora. We focus on three important types of non-task related dialogue acts: feedback, turn management and discourse structuring acts. We discussed in detail how single and multiple functions in these dimensions are expressed in different types of dialogue units, what linguistic and nonverbal means dialogue participants use for these purposes, and what aspects of a participant's behaviour are perceived as signals of these intentions. We revealed relations between observable features of communicative behaviour in different modalities and the intended multiple functions of multimodal utterances in dialogue. We also identified the general role of nonverbal signals for multimodal behaviour analysis in series of explorative and experimental studies.

Chapter 7 investigates automatic incremental dialogue act understanding on the basis of observable features such as linguistic cues, properties of intonation, and dialogue history using a token-based approach to utterance interpretation. We combined local classifiers that operate on low-level utterance and context features with global classifiers that incorporate the outputs of local classifiers applied to previous and subsequent tokens. We showed that applying this approach results in excellent dialogue act recognition scores for unsegmented spoken dialogue.

Chapter 8 presents a context-driven approach to the semantic interpretation and the generation of dialogue acts. We specify a multidimensional context model and show how (multiple) dialogue acts correspond to (multiple) context update operations on this model. A formalization of dialogue act update effects is proposed. We discuss context update mechanisms and the communicative effects of the understanding of dialogue behaviour which are the basis for dialogue participants to react in a certain way. The context-based generation of dialogue acts is addressed as well as the selection of alternative admissible dialogue acts. We formulate se-

mantic, pragmatic and linguistic constraints on dialogue act combinations for various types of dialogue unit, as well as the ordering of candidate dialogue acts according to their relative importance in a given context. We show that considering dialogue acts from various dimensions simultaneously contributes not only to more accurate and adequate interpretation of a user's dialogue behaviour, but also to the generation of interactive behaviour that is more natural to humans and exploits the full potential of spoken and multimodal interaction.

Chapter 9 reviews the most important conclusions and insights obtained in the thesis, and draws general conclusions with respect to the approach that we have applied. We finally suggest some perspectives for future research and application development on the basis of our results.

Samenvatting

Het hoofddoel van dit proefschrift is om bij te dragen aan de ontwikkeling van gefundeerde computermodellen van dialoogvoering. Communicatief gedrag in natuurlijke gesproken en multimodale dialogen wordt zowel empirisch als analytisch onderzocht. De benadering die wij voorstellen combineert een multidimensionaal perspectief op communicatie met een gestructureerde representatie van dialoogcontext.

De belangrijkste bijdragen van dit proefschrift zijn: (1) een theoretisch en empirisch gefundeerde definitie van het begrip 'dimensie' in de analyse van dialooghandelingen, die een basis vormt voor de keuze van dimensies in multidimensionale taxonomieën van dialooghandelingen en annotatieschema's; (2) een empirisch analyse van de multifunctionaliteit van dialooguitingen op basis van corpusgegevens; (3) de identificatie en succesvolle toepassing van eigenschappen van menselijk nonverbaal gedrag in een aantal klassen van dialooghandelingen, zoals feedback handelingen, handelingen voor beurtwisseling, en handelingen voor de structurering van de dialoog; (4) de ontwikkeling van een op machine leren gebaseerde methode voor het incrementeel interpreteren van dialooguitingen, in het bijzonder van hun communicatieve functies; (5) een context-gedreven benadering voor de interpretatie en generatie van multifunctionele dialooghandelingen.

In dit proefschrift wordt aangetoond dat het systematisch toepassen van een multidimensionale benadering van communicatie leidt tot beter inzicht in menselijk dialooggedrag en betere computationele dialoogrmodellen. Een waaier van problemen kan worden opgelost wanneer een multidimensionale benadering van dialooganalyse gebruikt wordt, met scherp omlijnde concepten van 'dimensie' en 'dialoogcontext' die gebaseerd zijn op gedetailleerde empirische analyse van dialooggedrag. De verkregen inzichten in de aard, de vormen en de talige en nonverbale manifestaties van multifunctionaliteit openen nieuwe perspectieven voor het ontwerp van dialoogsystemen die multifunctionele uitingen kunnen begrijpen en produceren.

Hoofdstuk 3 defineert een begrip 'dimensie' dat een conceptuele, theoretische en empirische betekenis heeft niet alleen voor de annotatie van dialooghandelingen, maar ook voor het segmenteren van dialogen in functionele eenheden. Wij stellen criteria op om te beslissen (1) welke dimensies zouden moeten worden onderscheiden en waarom; en (2) hoe zij in een taxonomie kunnen worden georganiseerd. Wij laten zien hoe deze criteria in operationele tests kunnen worden omgezet voor het ondersteunen van deze beslissingen. De toepassing van de criteria heeft een empirische basis gelegd onder de tien dimensies in de DIT^{++} taxonomie van dialooghandelingen en heeft geleid tot de keuze van negen dimensies voor de ISO 24617-2 standaard voor dialoogannotatie.

Hoofdstuk 4 richt zich op het annoteren van dialogen met dialooghandelingsinformatie. Multi- en één-dimensionale benaderingen van deze taak worden besproken en vergeleken. Geconcludeerd wordt dat multidimensionale annotatieschema's subtiele theoretische en empirische verschillen tussen bepaalde concepten beter kunnen vangen, en gemakkelijker aangepast kunnen worden voor verschillende doeleinden en taakdomeinen. In tegenstelling tot wat vaak gedacht werd, kunnen multidimensionale annotatieschema's betrouwbaar en efficient door annotatoren worden toegepast. Wij laten zien dat de voorgestelde multidimensionale manier om dialogen te segmenteren in functionele eenheden ('functionele segmenten') diverse bekende problemen in dialooganannotatie oplost die te maken hebben met onderbrekingen in de spraak, min of meer gelijktijdig spreken, en discontinuïteit van de segmenten die voor analyse en annotatie relevant zijn. Wij tonen ook aan dat een multidimensionale segmen-

tatiewijze resulteert in een meer nauwkeurige analyse en in betere automatische classificatie van dialooghandelingen (Hoofdstuk 7).

Verschillende verbeteringen en uitbreidingen van bestaande annotatieschema's van dialooghandelingen worden voorgesteld. Een uitbreiding betreft relaties die kunnen optreden tussen verschillende soorten eenheden in een dialoog. In een studie van het voorkomen van dergelijke relaties in dialoogcorpora hebben wij vier soorten relaties gevonden: retorische relaties tussen dialooghandelingen of tussen hun semantische inhoud; feedback relaties; en functionele afhankelijkheden. Sommige van deze relaties kunnen grotere eenheden of groepen van deze eenheden aan elkaar koppelen. Aangetoond wordt dat de verschillende soorten relaties significant verschillen in hun bereik en in de afstand tussen de aan elkaar gekoppelde eenheden. Een metamodel voor de annotatie van dialooghandelingen wordt voorgesteld dat het ISO 24617-2 metamodel uitbreidt met de vier soorten relaties tussen de diverse mogelijke soorten eenheden. Een andere uitbreiding, inmiddels opgenomen in het ISO 24617-2 metamodel, betreft de representatie van dialooghandelingen die gepaard gaan met uitdrukkingen van onzekerheid, voorwaardelijkheid, of emoties met behulp van 'qualifiers' die aan een communicatieve functie kunnen worden gekoppeld.

In Hoofdstuk 5 onderzoeken wij de vormen van multifunctionaliteit die in natuurlijke dialogen voorkomen. De diverse functies die een uiting heeft kunnen vaak gerelateerd worden aan verschillende aspecten van communicatie ('dimensies'). De relaties tussen vormen van multifunctionaliteit en onderscheiden communicatieve dimensies worden in detail bestudeerd. Wij doen dit analytisch door logische en pragmatische relaties tussen communicatieve functies te bestuderen en empirisch door de multifunctionaliteit te analyseren die in verschillende soorten dialoogeenheden voorkomt, zoals in een 'gewoon' functioneel segment, in een ingebed functioneel segment, of in een sequentie van functionele segmenten. De resultaten van deze studie hebben gevolgen voor de semantische interpretatie van dialooguitingen en voor hun generatie door een dialoogsysteem.

Hoofdstuk 6 heeft betrekking op de interpretatie van communicatief gedrag zoals in de geannoteerde dialoogcorpora waargenomen. Wij concentreren ons op drie belangrijke types van niet taak-gerichte dialooghandelingen: die voor feedback, voor beurtwisseling en voor dialoogstructurering. Wij bestuderen in detail hoe de communicatieve functies in deze dimensies in verschillende types van dialoogeenheden worden uitgedrukt, welke talige en nonverbale middelen hiervoor worden gebruikt, en welke aspecten van dialooggedrag worden waargenomen als signalen van deze intenties. Wij vonden interessante relaties tussen waarneembare eigenschappen van multimodale communicatieve uitingen en hun communicatieve functies, en identificeerden de rol van nonverbal signalen in multimodale communicatie in een aantal explorative en experimentele studies.

Hoofdstuk 7 onderzoekt de mogelijkheden van automatische incrementele interpretatie van dialooguitingen, d.w.z. interpretatie terwijl de uiting waargenomen wordt, op basis van waarneembare eigenschappen zoals talige kenmerken, prosodische eigenschappen, en dialooggeschiedenis. Wij laten zien dat een token-gebaseerde benadering waarin lokale classificatoren, die op low-level uiting- en contexteigenschappen werken, gecombineerd worden met globale classificatoren, die de output meenemen van lokale classificatoren toegepast op voorafgaande en volgende tokens, leidt tot uitstekende herkenningsscores van dialooghandelingen in ongesegmenteerde gesproken dialogen.

Hoofdstuk 8 stelt een context-gedreven benadering voor van de semantische interpretatie en de generatie van dialooghandelingen. Een multidimensionaal gestructureerd contextmodel wordt uitgewerkt, en gedemonstreerd wordt hoe (combinaties van) dialooghandelingen cor-

responderen met (combinaties van) updates van dit model. Een aantal mechanismen wordt beschreven die een rol spelen in updates van contextmodellen en die de basis vormen van het genereren van het vervolg van een dialoog. Een model van context-gebaseerde dialooggeneratie wordt besproken waarin in een eerste fase voor elke dimensie de contextueel mogelijke vervolg-dialooghandelingen gegenereerd worden, en in een volgende fase hieruit een selectie wordt gemaakt die uitgedrukt kan worden in een multifunctionele uiting. Semantische, pragmatische en linguïstische beperkingen op de combinaties van dialooghandelingen worden besproken, en de volgorde van alternatieve mogelijke dialooghandelingen. Deze benadering opent de mogelijkheid om interactief gedrag te genereren dat natuurlijk is voor menselijke dialoogpartners en de mogelijkheden van effectieve multimodale communicatie optimaal benut.

Hoofdstuk 9 vat de belangrijkste resultaten en inzichten samen die in dit proefschrift zijn verkregen, en trekt algemene conclusies met betrekking tot de benadering die wij hebben ontwikkeld en toegepast. Tenslotte schetsen wij enkele boeiende perspectieven voor toekomstig onderzoek en toepassingen op basis van onze resultaten.

TiCC dissertation series

1. Pashiera Barkhuysen. Audiovisual Prosody in Interaction. Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 3 October 2008.
2. Ben Torben-Nielsen. Dendritic morphology: function shapes structure. Promotores: H.J. van den Herik, E.O. Postma. Co-promotor: K.P. Tuyls. Tilburg, 3 December 2008.
3. Hans Stol. A framework for evidence-based policy making using IT. Promotor: H.J. van den Herik. Tilburg, 21 January 2009.
4. Jeroen Geertzen. Dialogue act recognition and prediction. Promotor: H. Bunt. Co-promotor: J.M.B. Terken. Tilburg, 11 February 2009.
5. Sander Canisius. Structured prediction for natural language processing. Promotores: A.P.J. van den Bosch, W. Daelemans. Tilburg, 13 February 2009.
6. Fritz Reul. New Architectures in Computer Chess. Promotor: H.J. van den Herik. Co-promotor: J.W.H.M. Uiterwijk. Tilburg, 17 June 2009.
7. Laurens van der Maaten. Feature Extraction from Visual Data. Promotores: E.O. Postma, H.J. van den Herik. Co-promotor: A.G. Lange. Tilburg, 23 June 2009 (cum laude).
8. Stephan Raaijmakers. Multinomial Language Learning. Promotores: W. Daelemans, A.P.J. van den Bosch. Tilburg, 1 December 2009.
9. Igor Berezhnoy. Digital Analysis of Paintings. Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 7 December 2009.
10. Toine Bogers. Recommender Systems for Social Bookmarking. Promotor: A.P.J. van den Bosch. Tilburg, 8 December 2009.
11. Sander Bakkes. Rapid Adaptation of Video Game AI. Promotor: H.J. van den Herik. Co-promotor: P. Spronck. Tilburg, 3 March 2010.
12. Maria Mos. Complex Lexical Items. Promotor: A.P.J. van den Bosch. Co-promotores: Dr. A. Vermeer, Dr. A. Backus. Tilburg, 12 May 2010 (in collaboration with the Department of Language and Culture Studies).
13. Marieke van Erp: Accessing Natural History. Discoveries in data cleaning, structuring, and retrieval. Promotor: A.P.J. van den Bosch. Tilburg, 30 June 2010.
14. Edwin Commandeur: Implicit Causality and Implicit Consequentiality in Language Comprehension. Promotores: Prof. dr. L.G.M. Noordman, Prof. dr. W. Vonk. Co-promotor: Dr. R. Cozijn. Tilburg, 30 June 2010.
15. Bart Bogaert: Cloud Content Contention. Promotores: Prof. dr. H.J. van den Herik, Prof. dr. E.O. Postma. Tilburg, 30 March 2011.
16. Xiaoyu Mao: Airport under Control. Promotor: Prof. dr. H.J. van den Herik, Prof. dr. E.O. Postma. Co-promotores: Dr. N. Roos and Dr. A. Salden. Tilburg, 25 May 2011.
17. Volha Petukhova: Multidimensional Dialogue Modelling. Promotor: Prof.dr. H. Bunt. Tilburg, 1 September 2011.